基于表征学习的离线强化学习方法研究综述

王雪松 [1] 王荣荣 [1] 程玉虎 [1]

摘 要 强化学习 (Reinforcement learning, RL) 通过智能体与环境在线交互来学习最优策略, 近年来已成为解决复杂环境下感知决策问题的重要手段. 然而, 在线收集数据的方式可能会引发安全、时间或成本等问题, 极大限制了强化学习在实际中的应用. 与此同时, 原始数据的维度高且结构复杂, 解决复杂高维数据输入问题也是强化学习面临的一大挑战. 幸运的是, 基于表征学习的离线强化学习能够仅从历史经验数据中学习策略, 而无需与环境产生交互. 它利用表征学习技术将离线数据集中的特征表示为低维向量, 然后利用这些向量来训练离线强化学习模型. 这种数据驱动的方式为实现通用人工智能提供了新契机. 为此, 对近期基于表征学习的离线强化学习方法进行全面综述. 首先给出离线强化学习的形式化描述, 然后从方法、基准数据集、离线策略评估与超参数选择 3 个层面对现有技术进行归纳整理, 进一步介绍离线强化学习在工业、推荐系统、智能驾驶等领域中的研究动态. 最后, 对全文进行总结, 并探讨基于表征学习的离线强化学习未来所面临的关键挑战与发展趋势, 以期为后续的研究提供有益参考.

关键词 强化学习, 离线强化学习, 表征学习, 历史经验数据, 分布偏移

引用格式 王雪松, 王荣荣, 程玉虎. 基于表征学习的离线强化学习方法研究综述. 自动化学报, 2024, **50**(6): 1104−1128

DOI 10.16383/j.aas.c230546

A Review of Offline Reinforcement Learning Based on Representation Learning

WANG Xue-Song[1] WANG Rong-Rong[1] CHENG Yu-Hu[1]

Abstract Reinforcement learning (RL), learning an optimal policy through online interaction between an agent and environment, has recently become an important tool to solve perceptual decision-making issues in complex environments. However, the online data collection may raise issues of security, time, or cost, greatly limiting the practical applications of reinforcement learning. Meanwhile, tackling intricate high-dimensional data input problems has also become a significant challenge for reinforcement learning due to the intricate and multifaceted nature of raw data. Fortunately, offline reinforcement learning based on representation learning can learn the policy only from historical experience data without interacting with the environment. It utilizes representation learning techniques to map the features of the offline dataset into low-dimensional vectors, which are subsequently employed to train the offline reinforcement learning model. This data-driven paradigm provides a new opportunity to realize the general artificial intelligence. To this end, this paper comprehensively reviews the recent research on offline reinforcement learning based on representation learning. Firstly, the problem setup of offline reinforcement learning is given. Then, the existing technologies are summarized from three aspects: Methodologies, benchmarks, offline policy evaluation and hyperparameter selection. Moreover, the study trends of offline reinforcement learning in industries, recommendation systems, intelligent driving, and other fields are introduced. Finally, the conclusion is drawn and the key challenges and development trends of offline reinforcement learning based on representation learning in the future are discussed, so as to provide a valuable reference for subsequent study.

Key words Reinforcement learning (RL), offline reinforcement learning, representation learning, historical experience data, distribution shift

Citation Wang Xue-Song, Wang Rong-Rong, Cheng Yu-Hu. A review of offline reinforcement learning based on representation learning. *Acta Automatica Sinica*, 2024, **50**(6): 1104−1128

强化学习 (Reinforcement learning, RL) 作为机器学习领域的一大重要分支, 近年来在各种复杂的决策控制任务中都发挥了重要作用[1−2]. 2016 年, DeepMind 公司创新性地将强化学习与系统神经科学相结合, 研发出 AlphaGo[3] 用于博弈游戏. 该程序成功击败了世界围棋高手李世石, 开创了深度强

收稿日期 2023-09-04 录用日期 2023-11-09
Manuscript received September 4, 2023; accepted November 9, 2023
国家自然科学基金 (62373364, 62176259), 江苏省重点研发计划项目 (BE2022095) 资助
Supported by National Natural Science Foundation of China (62373364, 62176259) and Key Research and Development Program of Jiangsu Province (BE2022095)
本文责任编委 杨涛
Recommended by Associate Editor YANG Tao
1. 中国矿业大学信息与控制工程学院 徐州 221116
1. School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116

化学习研究的先河. 随后, 针对不同应用场景, 该公司还研发了各种先进算法, 解决了许多领域中的关键科学问题, 例如: 用于 Atari 视频游戏的 MuZero[4]、用于生命科学领域解析蛋白质结构的 AlphaFold[5]、用于实现竞赛代码编程的 AlphaCode[6]、用于物理领域控制核聚变反应[7] 以及用于数学领域快速矩阵相乘的 AlphaTensor[8]. 经过近几年的不断发展与完善, 深度强化学习已然成为一大重要的决策工具. 然而, 现有的许多强化学习算法在仿真环境中能取得很好的效果, 但却难以用于真实业务场景. 其中一个制约因素在于智能体需要与环境进行大量交互, 一个高效的模拟器可能需要使用数以万计甚至数以亿计条轨迹并通过不断试错的方式来学习最优策略. 而在实际应用中, 主动在线交互可能导致智能体探索成本高、数据收集风险大且耗时长, 甚至引发巨大灾难.

幸运的是, 许多应用领域在前期已积攒大量历史经验数据, 如自动驾驶领域人类的行车记录[9] 与医疗领域患者的治疗记录[10] 等. 如何从这些固定的数据集中发现有价值的信息, 通过数据重用提高样本效率, 从而推断策略为用户提供安全决策支持, 是强化学习领域的重要研究课题. 为此, 离线强化学习[11−12] 应运而生. 与在线方式不同, 离线强化学习要求仅从固定的数据集中学习策略, 而无需与环境交互[11], 这种数据驱动的强化学习范式为研究从模拟环境到真实世界的转变提供了极大的可能. 然而, 想要从离线数据集中学到一个好的策略并非易事, 其中一大挑战在于, 智能体学习策略完全依赖于静态数据集, 而无法通过探索发现高奖励的状态–动作对, 另一关键挑战在于, 离线训练数据集与待测试的目标任务数据分布未必一致, 当行为策略与目标策略分布不同时, 会造成很严重的分布偏移问题[11]. 同时, 由于离线数据通常具有复杂且高维的特点, 传统的强化学习方法在处理这类数据时面临着巨大挑战.

为应对上述挑战, 近年来, 学者们对基于表征学习的离线强化学习方法展开深入研究. 表征学习是一种通过学习数据的内在特征来表示数据的机器学习方法. 当面对离线数据复杂且高维的大规模问题时, 有效地利用数据转换 (即表征学习), 通常可以显著提高离线强化学习过程的样本和计算效率. 研究证明, 利用在监督或无监督学习环境中开发的表征学习技术能够帮助智能体更有效地理解环境状态, 从而更快地找到最优决策策略. 因此, 基于表征学习的离线强化学习方法成为一个重要的研究方向.

具体而言, 基于表征学习的离线强化学习总体框架如图 1 所示, 其涵盖 4 个阶段: 数据收集、离线训练、策略选择和在线部署. 1) 在数据收集阶段, 智能体与环境进行交互, 通过执行一系列动作来收集训练数据, 以便为离线强化学习算法提供足够的训练数据. 这些数据包括智能体在不同状态下的观测值、采取的动作、与环境的交互结果以及相应的奖励信号. 2) 在离线训练阶段, 利用收集到的数据, 构建基于表征学习的离线强化学习模型. 首先, 将离线数据集中的原始动作、状态、轨迹、环境或任务等映射为潜在表征, 以揭示数据的内在结构和规律. 然后将潜在表征输入到离线强化学习模型中进行离线训练, 使得智能体能够更好地理解环境并做出合适的决策. 3) 在策略选择阶段, 需要根据离线训练得到的强化学习模型来选择最优策略, 以便在实际应用中实现最佳性能. 这个阶段通常涉及到评估和比较不同参数下的强化学习算法和策略, 以找到最适合特定任务的解决方案. 4) 在在线部署阶段, 智能体使用已经训练好的模型与策略部署到实际环境中. 智能体根据当前的观测值通过模型来预测最佳动作, 并执行该动作与真实环境进行交互. 这个阶

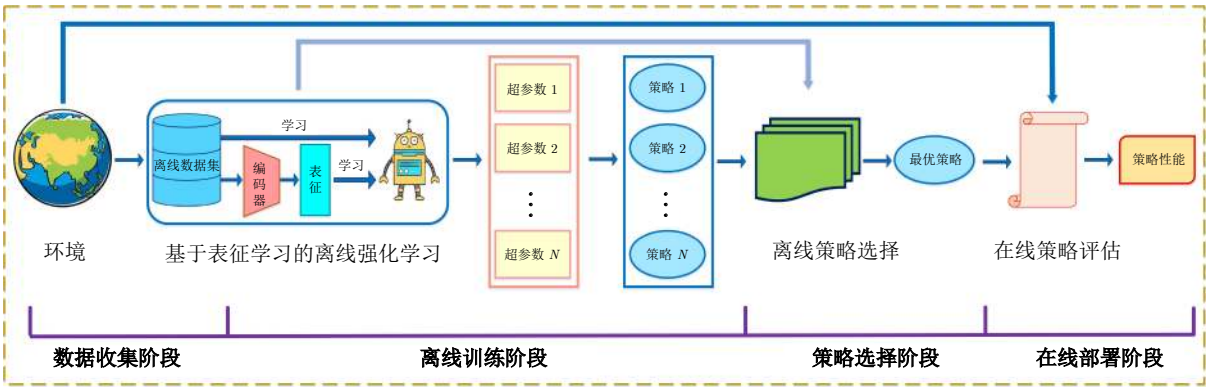


图 1　基于表征学习的离线强化学习总体框架

Fig. 1　The overall framework of offline reinforcement learning based on representation learning

段是将训练过程应用到实际环境中的关键环节. 整个基于表征学习的离线强化学习框架提供了一种强大而灵活的方法来解决强化学习领域中的问题. 通过离线训练和表征学习技术的结合, 智能体能够从历史数据中学习到更好的策略, 并在在线部署中取得更好的性能.

针对各阶段存在的问题, 本文从以下 4 个方面对目前的基于表征学习的离线强化学习方法进行综述与总结: 1) 在方法层面, 将现有的基于表征学习的离线强化学习方法归纳为动作表征、状态表征、状态−动作对表征、轨迹表征和任务或环境表征五大类; 2) 在数据层面, 详细介绍了 3 种离线强化学习基准数据集 RL Unplugged、D4RL、NeoRL 及其离线数据的构造方式; 3) 在评估层面, 总结现有的离线策略评估与超参数选择方法; 4) 在应用层面, 介绍离线强化学习在工业、推荐系统、智能驾驶等领域的应用. 最后, 给出结论与展望, 希望能为离线强化学习的研究人员提供参考.

1 问题描述

在在线强化学习中, 智能体与环境的交互可以通过马尔科夫决策过程来表示: $M=\{S,\ A,\ r,\ T,\ \gamma,\ \rho\}$, 其中 S 与 A 分别表示状态与动作空间, r 为奖励函数, $T(s_{t+1}|s_t,\ a_t)$ 为给定 t 时刻状态−动作对的状态转移函数, γ 为折扣因子, $\rho(s_0)$ 表示初始状态分布. 通过最大化期望累积折扣奖励 $R=\mathrm{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,\ \pi(a_t|s_t))\right]$, 从而学到最优策略 π^*. $Q(s_t,\ a_t)$ 为状态−动作值函数, 表示在给定状态 s_t 下进行动作 a_t 并执行策略 π 得到的期望总回报. 在策略评估步骤中, 通过最小化贝尔曼误差来获得 Q 函数的迭代更新式为

$$Q^{(i+1)}\leftarrow \arg\min_{Q}\mathrm{E}\left[r_t+Q^{(i)}(s_{t+1},a_{t+1})-Q^{(i)}(s_t,a_t)\right] \tag{1}$$

其中, i 表示迭代次数.

在策略提升步骤中, 对策略进行训练, 使在给定状态 s_t 执行动作 a_t 的状态−动作值最大化, 即

$$\pi^{(i+1)}\leftarrow \arg\max_{\pi}\mathrm{E}\left[Q(s_t,\pi^{(i)}(a_t|s_t))\right] \tag{2}$$

然而, 在离线强化学习设定中, 智能体无法与环境进行交互, 只能从由行为策略收集得到的固定数据集 D 中学习目标策略 π. 因此, 策略提升步骤的更新式为

$$\pi^{(i+1)}\leftarrow \arg\max_{\pi}\mathrm{E}_{s_t\sim D}\left[Q(s_t,\pi^{(i)}(a_t|s_t))\right] \tag{3}$$

与在线强化学习相比, 离线强化学习中智能体只能从固定的离线数据集中学习最优策略. 由于智能体无法与环境交互, 因此在离线强化学习设置下直接使用异策略在线强化学习方法会产生严重的外推误差[11], 从而导致智能体无法学习到最优策略. 造成外推误差的两种最主要原因为离线数据覆盖不足与分布偏移. 针对离线数据覆盖不足的问题, 很多离线强化学习算法事先假设离线数据集能充分覆盖到高奖励区域或者对该假设条件进行弱化, 抑或使用数据增强来扩充离线数据集. 而针对分布偏移, 近年来学者们提出很多解决方案, 使用生成模型对行为策略进行建模, 根据状态重构得到目标策略, 通过约束目标策略的分布, 使之与行为策略分布相接近, 从而避免产生分布外动作[13−14]. 除此之外, 复杂的高维数据输入是阻碍离线强化学习在真实系统中成功应用的一个难点, 特别是连续动作空间上的视觉观测数据输入问题. 表征学习技术可以将离线数据中的特征表示为低维向量, 是提高离线强化学习方法性能的关键. 为此, 在第 2 节中, 本文将详细阐述现有的基于表征学习的离线强化学习方法, 并探讨它们的优缺点和适用场景.

2 方法分类

本文使用关键词“离线强化学习”、“表征学习”、“编码器”及其相关组合, 在谷歌学术、Web of Science 和中国知网等平台上检索了 2019 年以来基于表征学习的离线强化学习算法. 通过筛选, 本文选择来自一流期刊或会议、具有较高引用率以及与讨论主题具有强相关性的文献. 在此基础上, 本文对现有的基于表征学习的离线强化学习方法进行分类. 根据强化学习的表征对象, 将这些方法分为五类: 动作表征、状态表征、状态−动作对表征、轨迹表征以及任务或环境表征.

1) 动作表征. 在已知状态的条件下, 使用编码器−解码器结构来得到动作表征. 其中, 编码器将动作映射到低维潜在空间, 解码器则从潜在空间预测得到动作. 这种方法能够加快学习最优策略的速度.

2) 状态表征. 通过学习状态的低维表示或将状态映射到特征空间来降低状态空间的维度. 这种方法能够提高学习的效率和泛化性能.

3) 状态−动作对表征. 将状态和动作信息一起编码, 同时考虑它们之间的相互关系. 通过生成更丰富的表征, 可以更好地捕捉到环境的状态和动作之间的复杂关系, 进一步提高强化学习的性能.

4) 轨迹表征. 将强化学习视为序列生成问题, 通过序列建模学习过去多个时间步的行为序列来预

测未来的动作. 这种方式可以捕捉到序列中的长期依赖关系, 并更好地表示问题.

5) 任务或环境表征. 借助元学习的思想, 学习任务或环境信息的内在特征. 这种方法可以帮助智能体更好地适应不同的任务或环境变化, 提高智能体的泛化能力.

以上五类表征学习方法在离线强化学习中发挥着重要作用, 它们在不同场景下都有各自独特的优势和适用性. 这些方法为解决离线强化学习中的复杂问题提供了有力的工具和思路, 基于表征学习的离线强化学习方法对比如表 1 所示, 下面将对这五类方法进行详细综述.

2.1　动作表征

由于强化学习中策略空间的高度复杂性, 直接在该空间中优化策略是困难的. 动作表征是指将原先的高维动作空间映射到低维表征空间, 并在该表征空间中对策略进行优化. 通过这种方式, 可以捕捉到动作的内在特征. 基于动作表征的离线强化学习方法通常使用编码器–解码器架构, 如条件变分自编码器 (Conditional variational auto-encoder, CVAE)[15−21]、流模型[22−23] 和扩散模型[24−25]. 在给定状态的情况下, 使用编码器–解码器架构获得动作表征, 从而大大降低在策略空间中寻找最优策略的难度. 同时, 这种技术将目标策略限制在行为策略的数据支持范围内, 有助于缓解分布偏移问题.

本文以编码器结构为例, 以 Actor-Critic 为强化学习主体网络结构, 参考文献 [52] 绘制了基于动作表征的离线强化学习框架, 如图 2 所示. 其中, 离线数据集中的动作经过编码器得到动作表征 z, 策略 $\pi_\theta(a|s)$ 给出智能体在给定环境状态 s 的条件下选择动作 a 的概率, θ 为 Actor 网络参数. 状态动作值函数 $Q_\psi(s,\ z)$ 用于评估在状态 s 的条件下执行动作表征 z 的价值, ψ 为 Critic 网络参数. L_{Actor} 与 L_{Critic} 分别表示策略网络 Actor 与价值网络 Critic 的损失函数.

变分自编码器 (Variational auto-encoder, VAE) 是由 Kingma 等[53] 于 2014 年提出的基于贝叶斯推断的网络模型, 具有较好的数据生成效果和可解释性. 为此, Fujimoto 等[15] 于 2019 年率先提出批约束深度 Q 学习 (Batch-constrained deep Q-learning, BCQ), 利用变分自编码器来生成与离线数据集分布相近的动作, 并结合一个扰动模型对生成的动作进行调优以使动作具有多样性. 测试阶段, 仅在生成的动作空间中选择使 Q 值最大的那些动作而不考虑分布外动作. BCQ 不涉及对未知状态–动作对的考虑, 因此不会在策略与值函数上引入额外的偏差; 同时, 动作与值函数分开学习, 也在一定程度上避免了误差累积. Fujimoto 等[15] 在连续控制任务上验证了 BCQ 的值估计准确且稳定, 说明外推误差得到一定程度的缓解. 然而, 在某些情况下, 严格限制目标策略近似行为策略的方法并不一定适用. 特别是在离线样本数量不足的情况下, BCQ 算法可能会受到行为策略分布密度的制约, 模型将难

表 1　基于表征学习的离线强化学习方法对比

Table 1　Comparison of offline reinforcement learning based on representation learning

表征对象	参考文献	表征网络架构	环境建模方式	应用场景	特点	缺点
动作表征	[15−21]	VAE	无模型	机器人控制、导航	状态条件下生成动作, 将目标策略限制在行为策略范围内, 缓解分布偏移	不适用于离散动作空间
	[22−23]	流模型				
	[24−25]	扩散模型				
状态表征	[26−27]	VAE	无模型	基于视觉的机器人控制	压缩高维观测状态, 减少冗余信息, 提高泛化能力	限定于图像 (像素) 输入
	[28]	VAE	基于模型			
	[29]	GAN	基于模型			
	[30]	编码器架构	基于模型			
	[31−32]	编码器架构	无模型			
状态–动作对表征	[33]	自编码器	基于模型	基于视觉的机器人控制、游戏、自动驾驶	学习状态–动作联合表征, 捕捉两者交互关系, 指导后续决策任务	限定于图像 (像素) 输入
	[34]	VAE	基于模型			
	[35−36]	编码器架构	无模型			
	[37−38]	编码器架构	基于模型			
轨迹表征	[39−44]	Transformer	序列模型	机器人控制、导航、游戏	将强化学习视为条件序列建模问题, 用于预测未来轨迹序列	轨迹生成速度慢, 调优成本高
	[45−47]	扩散模型				
任务表征	[48−49]	编码器架构	无模型	机器人控制、导航	借助元学习思想, 使智能体快速适应新任务	泛化能力依赖于任务或环境之间的相似性
环境表征	[50−51]	编码器架构	基于模型			

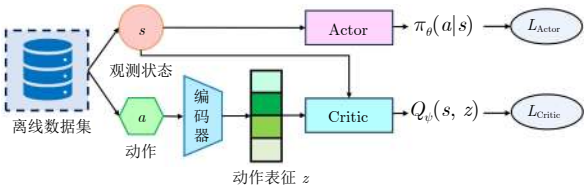


图 2 基于动作表征的离线强化学习框架

Fig. 2 The framework of offline reinforcement learning based on action representation

以准确地反映出未知分布的特征.

He 等[16] 从分布式强化学习角度出发, 提出一种悲观离线策略优化 (Pessimistic offline policy optimization, POPO), 其主要由变分自编码器和悲观分布式 Critic 构成. POPO 的主要思路为: 首先, Critic 采用分位数回归法学习一种具有风险规避偏好的 Q 值函数分布, 以获得悲观值函数来控制真实值函数与估计值函数的估计偏差; 然后, Actor 直接采用变分自编码器将目标策略约束在行为策略附近, 通过最大化 Critic 估计的 Q 值函数来显式地输出动作. POPO 通过设计保守或悲观值函数来阻止策略访问分布外动作, 在一些相关任务上能够获得优越性能. 但是, 这种过于严格地限制访问分布外动作的方法往往只能学到保守的次优策略, 降低了神经网络在分布外区域的泛化性能.

Wu 等[17] 考虑了限制性较小的支持约束, 而不是条件过于严苛的密度策略约束, 通过显式密度估计将目标策略保持在行为策略的支持范围内, 因此提出支持策略优化方法 SPOT (Supported policy optimization). 该方法采用 CVAE, 对行为策略的支持集进行显式建模, 从而可以直接计算散度, 并提出了一个简单而有效的基于密度的正则化项, 该正则化项可以插入到现有的异策略强化学习算法中. 然而, 在在线阶段, SPOT 逐步减小训练目标的保守性, 而不是保持相同的离线训练目标. 文献 [54] 指出这种方式并没有显著改进所考虑的基准任务在探索方面的性能.

为提升离线强化学习中值函数网络的泛化性, 相对地从源头修改贝尔曼算子可能更为有效. 为此, Lyu 等[18] 提出轻度保守 Q 学习 (Mildly conservative Q-learning, MCQ) 来积极训练分布外动作, 以缓解现有离线强化学习方法中的过度悲观. MCQ 的核心是轻度保守贝尔曼算子的设计, 具体分为两个步骤: 首先将分布外动作的值函数全部赋成比离线数据集中动作的最大 Q 值小一点的值, 然后再进行正常的一步贝尔曼更新. 尽管 MCQ 利用 CVAE 建模行为策略仍然可能产生分布外的动作并查询未定义的 Q 值, 但由于轻度保守贝尔曼算子的保证, 分布外动作的值函数高估误差实际是可控的.

类似地, Rezaeifar 等[19] 从强化学习探索的角度来考虑值函数学习的稳定性, 将离线强化学习表述为“反探索”问题. 受基于红利的探索方法启发, Rezaeifar 等[19] 将在离线状态–动作对上训练得到的变分自编码器的重构误差作为探索红利, 从奖励中减去探索红利而不是增加探索红利, 从而使得所求策略能够近似离线数据集的分布, 避免了智能体探索未知的动作. 但需要注意的是, 在部署实施这种反探索方法的过程中需要使奖励范围在不同环境中保持一致, 否则会改变最优策略集.

先前的方法将策略约束作为正则化项纳入到策略或 Q 值函数的优化过程中, 从而避免产生分布外动作. 这类方法需要引入额外的正则化因子来权衡原始目标与约束条件, 但在实际应用中很难选择到合适的正则化因子. 此外, 考虑到具有多样性动作的数据集, 距离度量的选择 (如 KL (Kullback-Leibler) 散度与最大均值差异) 可能限制过于严格, 进而导致策略退化为数据集上的行为克隆. 为此, Zhou 等[20] 提出一种隐动作空间内策略 (Policy in the latent action space, PLAS) 学习方法, 以隐式的方式对策略进行约束, 使智能体在数据集支撑范围内选取动作. 具体而言, Zhou 等[20] 将行为策略建模为 CVAE, 在 CVAE 的隐动作空间中学习隐策略, 从而达到约束策略的目的, 再使用解码器在环境的原始动作空间中输出动作. 这种约束的优势在于: PLAS 能够隐式地将策略限制在数据集的支撑范围内, 其不会受到行为策略分布密度的限制, 也不会影响其他变量的优化.

进一步, Chen 等[21] 将 PLAS 与优势加权回归算法相结合, 提出一种隐变量优势加权策略优化 (Latent-variable advantage-weighted policy optimization, LAPO) 方法, 主要由一个优势加权的行为策略模型和一个隐策略模型构成. 在 LAPO 中, Chen 等[21] 采用优势加权回归的思想来最大化优势加权动作的对数似然, 设计了一个变分自编码器来重建产生多模态数据的行为策略. D4RL 基准数据集上的评估结果表明: LAPO 展示出在多模态离线强化学习任务上的优势. 然而, 与 BCQ 类似, PLAS 和 LAPO 也是使用预先训练好的变分自编码器来近似整个离线数据集的分布, 且该分布在训练隐策略时是固定不变的. 因此当离线数据量较小时, 模型对高回报样本的表达能力可能会受到限制.

PLAS 对变分自编码器的特定使用导致必须要裁剪隐空间: 将隐空间中策略输出限制在一个固定的范围内, 否则其训练过程不稳定. 尽管裁剪过程是有效的, 但这种人为手动裁剪的方式易于裁剪掉

分布内的动作, 从而限制离线强化学习方法的性能. 此外, 变分自编码器采用似然函数的变分下界代替真实的数据分布, 只能得到真实数据的近似分布. 相较于变分自编码器, 另一种重要的深度生成模型——流模型具有准确的似然估计, 生成的样本更接近真实数据曲线[55]. 为此, Akimov 等[22] 利用保守标准化流 (Conservative normalizing flow, CNF) 来为离线强化学习创建一个更好的动作编码器模型, 通过在标准化流模型的最后一层添加可逆 Tanh 激活函数来使得动作编码器可以利用整个隐动作空间, 一方面避免了事后隐空间裁剪, 另一方面也避免了在分布外生成动作.

Yang 等[23] 研究离线分层强化学习并给出由原始误差、离线误差和表征误差构成的次优性能下界, 其中, 原始误差来源于通过学习得到的下层策略与其真值之间的泛化误差, 离线误差是指在高层离线数据集中学习产生的误差, 表征误差则是由分层结构的有限表征能力导致的. 鉴于以往方法对策略空间的表征能力不足, Yang 等[23] 提出一种无损性能下层策略发现 (Lossless primitive discovery, LPD), 主要思路为: 采用标准化流模型学习策略表征, 利用可逆函数将隐向量映射到临时扩展的动作上. 由于该映射是可逆的, LPD 能够确保智能体对原始策略空间进行完整的恢复, 进而提高分层策略的性能. 通过分析发现: 上述基于流模型的离线强化学习使用的是确定性轨迹, 原始数据空间与潜在空间需要通过可逆映射函数进行连接且要求该映射函数的雅克比行列式能够计算. 这些约束条件直接限制了映射函数的选取, 进而限制了流模型的策略表达能力.

生成式建模的一个核心问题是模型的灵活性和可计算性之间的权衡. 扩散模型[56] 的基本思想是利用正向扩散过程来系统地扰动数据中的分布, 然后通过学习反向扩散过程恢复数据的分布, 这样就产生一个高度灵活且易于计算的生成模型. 扩散模型已成为目前性能最为优越的深度生成模型, 其在计算机视觉和语音生成等领域取得巨大的成功. 最近, 陆续有学者开始尝试利用扩散模型来解决离线强化学习的分布偏移问题.

为提高目标策略的表达能力, Wang 等[24] 构造一个基于多层感知机的“以状态为条件–以动作为输出”的条件扩散模型并利用其来生成策略, 提出一种扩散 Q 学习 (Diffusion-QL) 方法. 条件扩散模型的目标函数由行为克隆项和策略改进项构成, 其中, 行为克隆项为策略正则化损失, 用于鼓励扩散模型生成与离线数据集分布相一致的动作; 策略改进项为 Q 值函数正则化项, 根据学到的 Q 值来对高价值动作进行采样. D4RL 基准数据集上的评估结果表明: 扩散 Q 学习能够很好地捕获多模态分布, 提高策略的表达能力.

Diffusion-QL 将扩散模型用于离线强化学习来提高策略表达能力, 然而该方法使用扩散模型作为最终策略的隐式正则化项, 而不是显式策略先验. 相比之下, Chen 等[25] 更侧重于使用扩散模型进行单步决策, 认为有限的策略表达能力是引起外推误差的主要原因, 为此采用扩散模型对多样性策略进行“高保真度” (High-fidelity) 的建模, 提出一种从行为策略候选集中选择动作的离线强化学习方法 SfBC (Selecting from behavior candidates). 具体而言, Chen 等[25] 将目标策略分解为生成式行为模型和动作评估模型两个部分, 利用重要性采样技术从生成式行为模型中采样候选动作, 并通过动作评估模型计算候选动作的重要性权重. 仿真结果表明, SfBC 能提升模型分布的表达能力, 减少离线强化学习中的外推误差, 进而提升学习性能.

基于动作表征的离线强化学习通过编码器–解码器架构 (如 CVAE、流模型、扩散模型) 生成状态条件下的动作, 以限制目标策略在行为策略范围内. 这种限制有助于防止策略偏离目标太远, 从而减轻分布偏移问题. 由于其能够有效地处理具有复杂动态的连续动作空间中的问题, 此类算法在机器人连续控制任务中得到广泛应用, 提高了机器人在物体抓取、姿势控制等任务中的执行精度. 在工业生产线等实际应用中, 机器人需要执行一系列复杂的操作, 如搬运、装配和焊接等. 通过使用基于动作表征的离线强化学习算法, 机器人可以在不需要人工干预的情况下自主学习和优化其操作策略, 从而提高生产效率和降低生产成本. 此外, 这类方法在推荐系统[57]、智能驾驶[58]、能源管理[59] 等领域也取得了良好效果. 然而, 对于离散动作场景, 其动作空间中的动作通常是有限且已知的, 使用基于动作表征的离线强化学习算法难以捕捉有效的特征表示, 需要选择其他适用的算法.

2.2 状态表征

在视觉强化学习领域, 环境的观测状态通常以图像或像素的形式表示. 然而, 由于像素输入的维度通常非常高, 直接使用这些像素作为输入会增加计算和学习的复杂性. 状态表征是指将输入的高维状态空间压缩为低维的状态表征. 它可以从原始观测状态中提取出关键信息, 减少冗余信息. 在实践中, 常见的方法是利用变分自编码器[26–28]、生成对抗网络 (Generative adversarial network, GAN)[29] 等

生成模型来生成状态, 或者使用编码器架构[30−32] 来学习潜在的状态表征. 通过这些方式, 数据在低维空间中更加易于处理和分析, 从而提高学习的效率和泛化性能.

同样以编码器结构为例, 以 Actor-Critic 为强化学习主体网络结构, 参考文献 [32] 绘制了基于状态表征的离线强化学习框架, 如图 3 所示. 其中, 离线数据集中的观测状态 (通常为图像) 经过编码器得到状态表征 z, 策略 $\pi_\theta(a|z)$ 给出智能体在给定状态表征 z 的条件下选择动作 a 的概率, θ 为 Actor 网络参数. 状态动作值函数 $Q_\psi(z,\ a)$ 用于评估在状态表征 z 的条件下执行动作 a 的价值, ψ 为 Critic 网络参数. L_{Actor} 与 L_{Critic} 分别表示策略网络 Actor 与价值网络 Critic 的损失函数.

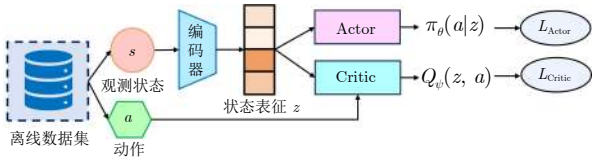


图 3 基于状态表征的离线强化学习框架

Fig. 3 The framework of offline reinforcement learning based on state representation

离线强化学习的目的是在固定的离线数据集中实现预期累积奖励的最大化. 目前离线强化学习方法的基本原则是将策略限制在离线数据集的动作空间中. 然而, 这些方法忽略了数据集的轨迹无法完全覆盖状态空间的情况. 为此, Zhang 等[26] 提出状态偏差校正 (State deviation correction, SDC) 算法来减少训练策略和离线数据集之间的状态访问不匹配问题. 其基本思想是预测执行策略的结果, 并确保生成的状态是在数据集的支持范围内. SDC 由 3 个部分组成: 动力学模型、状态转移模型和 Actor-Critic 智能体. 首先, 通过训练动力学模型, 根据当前状态与动作来预测下一个状态. 其次, 维护状态转移模型, 将当前状态作为输入并预测下一个状态. 该状态转移模型由 VAE 建模, 它规定了状态的支持范围, 与动作无关. 然后, 对策略进行训练, 使其能够在扰动初始状态时产生将智能体引向分布域内的动作, 并用动力学模型来预测下一个时间步骤的状态. 因此, 智能体有望预测潜在的下一个状态, 并逐渐靠近离线数据集的状态, 从而减少状态偏差. 仿真结果表明, SDC 优于现有的策略约束方法.

Weissenbacher 等[27] 提出 Koopman 前向 Q 学习 (Koopman forward conservative Q-learning, KFC) 算法来解决值函数的限制性泛化问题, 修改 Koopman 算子时关注点不再是动作, 而是状态. 具体而言, 通过学习 Koopman 潜在表示来推断系统潜在动力学的对称性, 利用这些环境动力学的对称性以自监督的方式训练 VAE 来扩展对环境状态空间的探索, 同时限制分布外的泛化误差, 以此来增强值函数泛化能力. 然而 Koopman 理论存在着两大局限性: 仅适用于具有可微状态转移的动力学系统, 且系统的 Koopman 算子采用一个双线性化模型, 因此所提方法无法处理非连续任务.

先前的离线强化学习方法主要侧重于压缩原始高维状态以获得紧凑的状态表征, 然而却忽略了直接从丰富多样的观测空间 (如图像) 中学习的能力. 为此, Rafailov 等[28] 提出一种基于模型的策略优化算法 LOMPO (Latent offline model-based policy optimization), 该方法能够有效处理高维图像输入, 利用潜在动力学模型对高维视觉观测空间进行建模, 并在潜在空间中表示不确定性, 从而辅助强化学习算法的决策过程. 具体而言, 首先将原始观测图像作为输入, 利用变分模型得到当前时刻的潜在状态表征, 再将该表征与当前时刻的动作共同输入到潜在动力学模型中, 预测得到下一时刻的潜在状态表征. 然后, 构建一个具有不确定性惩罚的潜在马尔科夫决策过程, 在奖励函数中引入不确定性惩罚项. 最后, 使用变分下界来学习最优策略. 然而, 该方法仅利用图像作为输入, 因此会导致重构图像中的信息缺失或动力学模型估计不准确.

离线强化学习在训练过程中无法与物理环境进行交互, 因此存在固有的分布偏移问题. 为解决这一问题, 先前的方法采取状态增广技术, 通过学习回放池的历史经验数据来建立动力学模型, 并利用生成的预测状态来扩充数据集. 为将这一优势应用到基于图像的强化学习领域, Cho 等[29] 提出一个从状态到像素的生成模型 S2P (State to pixel). 首先, 输入当前时刻的状态和相应的图像观测值, 通过基于多模态仿射变换的生成对抗网络合成出下一时刻的图像观测样本; 然后, 利用 L_1 范数计算生成图像与真实图像之间的像素级损失和感知相似度损失; 最后将生成的图像纳入到基于模型的离线强化学习算法中. 多模态仿射变换模块能够有效地利用状态和图像之间的跨模态信息, 从可靠的状态转换中生成动态一致的图像. 实验结果表明, 基于 S2P 的图像合成方法不仅提高了基于图像的离线强化学习性能, 而且在未知任务上也展现出强大的泛化能力.

为处理复杂高维输入下的稀疏奖励控制任务, Gieselmann 等[30] 将学习潜在空间模型与基于采样的规划技术相结合, 提出一种潜在空间树搜索规划算法 VELAP (Value-guided expansive latent

planning). 该算法将当前状态编码为潜在空间中的状态表征, 并将其与动作共同输入到潜在动力学模型中, 从而得到下一时刻的状态表征, 并利用对比学习来优化该潜在动力学模型. 此外, 该算法还使用局部策略与全局策略在潜在空间中实现启发式的稀疏树搜索, 以寻找使回报值最高的路径规划策略. 其中, 搜索树作为状态覆盖的存储器, 在规划过程中引导智能体朝向数据支持范围内未探索的区域. 最后, 利用条件生成模型 CVAE, 在给定状态条件下从动作分布中进行采样, 从而得到动作. 在高维的机器人操作任务中展现出 VELAP 的优越性.

离线强化学习算法通常面临两个主要问题: 一是从高维视觉输入数据中进行有效学习相当困难; 二是隐式欠参数化 (Implicit under-parameterization) 现象可能会导致值网络的低秩特征, 进而影响算法性能. 为解决上述问题, Zang 等[31] 将状态表征学习与离线策略训练过程分离开来, 并提出一种离线状态表征方法 BPR (Behavior prior representation). 该方法首先通过模仿数据集中的行为来隐式地学习状态表征, 并将其归一化到单位超球面上. 然后, 冻结学到的编码器, 并在固定状态表征的基础上使用现有的离线强化学习算法来训练策略. BPR 能够灵活地与现有的离线强化学习方法相结合使用. 然而, 这种模仿学习方式只是很好地拟合行为策略, 当测试环境的分布与训练环境分布差异很大时, BPR 仍然无法提升表征的泛化能力.

在线强化学习泛化算法的性能在离线设置中可能会受到限制, 这是由于离线数据集的大小和质量制约了智能体提高零样本 (Zero-shot) 泛化性能的能力, 尤其是在处理高维观测的离线数据情况下. Mazoure 等[32] 指出, 先前的方法对观测值之间的相似性估计不准确, 具有相似未来行为的观测样本应该被分配到相近的表征空间中. 基于这一假设, 他们提出广义相似性函数 (Generalized similarity functions, GSF), 用来计算相对于任何瞬时累积信号的观测样本对之间的相似性. 该方法使用自监督对比学习来训练离线强化学习智能体, 并通过比较观测样本的期望未来行为相似性来聚合其状态表征. 实验表明, GSF 可以在基于像素的控制任务中提高零样本泛化性能. 然而, 实验结果很大程度上依赖于超参数的选取.

基于状态表征的离线强化学习通过压缩高维观测状态, 将复杂的观测状态转化为更简单、易于处理的低维表示, 使模型更加关注与目标任务相关的关键信息, 同时忽略不相关或冗余的信息, 进而提高模型的学习效率与泛化能力. 这类方法在基于视觉感知的观测输入场景中尤为有效, 例如图像或像素数据. 在实际应用中, 基于状态表征的离线强化学习算法已经在许多领域取得显著的成果, 如基于视觉的机器人控制和推荐系统[60−61] 等. 然而, 此类算法也存在一些挑战: 如何设计合适的状态表征是一个关键问题, 过于简单可能无法充分利用高维观测状态的信息, 过于复杂则会增加计算复杂度并导致过拟合. 因此, 未来需要进一步提高算法的性能与应用范围.

2.3　状态−动作对表征

状态−动作对表征是指将状态和动作作为一个整体进行编码, 从而得到一个潜在的表征向量. 这个表征向量能够捕捉到状态和动作之间的交互关系, 并且可以用于下一步的决策过程. 常见的方法为利用编码器−解码器架构[33−34] 或者编码器架构[35−38] 学习状态−动作对的潜在表征, 然后将学到的联合表征用于指导后续的离线强化学习决策任务.

参考文献 [36], 基于状态−动作对表征的离线强化学习框架如图 4 所示. 其中, 将离线数据集中的观测状态 (通常为图像) 与当前时刻的动作作为一个整体, 经过编码器得到状态−动作对表征 $\phi_\xi(s,\ a)$, 其中, ξ 为编码器网络参数. 策略 $\pi_\theta(a|s)$ 给出了智能体在给定观测状态 s 的条件下选择动作 a 的概率, θ 为 Actor 网络参数. 状态动作值函数 $Q_\psi(s,\ a)$ 用于评估在状态 s 的条件下执行动作 a 的价值, ψ 为 Critic 网络参数. L_{Actor} 表示策略网络 Actor 的损失函数, 状态−动作对表征 $\phi_\xi(s,\ a)$ 与状态动作值函数 $Q_\psi(s,\ a)$ 共同构成价值网络 Critic 的损失函数 L_{Critic}.

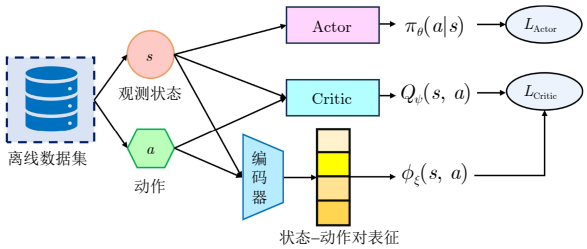


图 4　基于状态−动作对表征的离线强化学习框架

Fig. 4　The framework of offline reinforcement learning based on state-action pairs representation

在基于模型的离线强化学习研究中, 文献 [62] 表明选择不同的模型不确定性估计方法会对性能产生显著影响. 然而, 现有算法难以在实践中计算模型不确定性的距离, 导致估计不可靠. 因此, 目前仍未解决如何选择模型不确定性估计方法以及哪种保守性方法更适用于实际需求的问题. 为此, Kim 等[33] 提出基于计数的保守性离线强化学习方法 (Count-

based conservatism for model-based offline RL, Count-MORL). 该方法利用离线数据集中状态–动作对的估计频率 (计数) 来量化目标和真实过渡动力学模型之间的估计误差. 通过理论证明, 估计误差是有界的, 且与状态–动作对的频率成反比. 基于这一发现, Kim 等[33] 构造了一个基于计数的保守马尔科夫决策过程. 在该方法中, 利用自编码器来学习状态–动作对的潜在表征, 这些潜在表征再经过计数函数用于估计近似计数值. 在 D4RL 基准数据集上展示出 Count-MORL 相较于现有的基于模型的离线强化学习方法的优越性. 这些结果说明, 针对基于模型的离线强化学习, 采用基于计数的保守性策略不仅高效, 而且具有实际应用价值.

基于模型的离线强化学习方法通常依赖于对模型误差边界的精确估计. 这种估计通常是通过不确定性估计方法来实现的, 主要包括参数化与非参数化两种方式. 在处理大规模数据集时, 参数化方法效果较好, 然而, 如果模型设定不恰当, 可能会影响其准确性. 相反, 非参数化方法更适用于数据有限的情况, 但需要选择合适的度量方式以保证结果的有效性. Tennenholtz 等[34] 将两者相结合以进行不确定性估计并提出 GELATO (Geometrically enriched latent model for offline RL). 具体而言, 利用黎曼几何的最新成果, 在特征空间上建立置信区间, 通过测量新样本到数据流形的平均测地距离来估计模型误差, 而不是使用欧氏距离. 此外, 在已有的 VAE 模型基础上添加一个潜在的前向函数, 状态–动作表征分别通过奖励预测器和潜在前向模型来生成预测奖励与下一时刻的状态表征. 这种方法能够正确区分生成参数前向模型的认知不确定性 (数据缺失) 和偶然不确定性 (环境动力学). 然后, 构造奖励惩罚的马尔科夫决策过程, 并将误差作为悲观正则化器. 最后, 在悲观马尔科夫决策过程上训练强化学习智能体. 连续控制和自动驾驶基准测试结果表明, GELATO 能够有效地对离线强化学习智能体进行惩罚. 然而, 由于需要计算测地距离, 导致算法计算速度较慢.

在先前的离线强化学习工作中, 研究者们通常使用条件扩散模型来获得策略表达, 以反映数据集中的多模态行为. 然而, 这些方法并未充分考虑到分布外状态的泛化性能. Ada 等[35] 提出一种用于扩散策略的状态重构 (State reconstruction for diffusion policies, SRDP) 方法, 该方法在先前 Diffusion-QL 的基础上引入辅助状态重构特征学习来解决分布外泛化问题. 具体而言, SRDP 使用基于自编码器的状态重构损失作为辅助目标, 从离线强化学习数据集中提取更丰富的描述性特征, 从而指导扩散策略的学习过程. 实验结果表明, 与之前的方法相比, SRDP 能够更准确地表达多模态策略, 更好地划分状态空间, 并实现更快的收敛速度.

通常情况下, 在强化学习实际部署前需要进行异策略评估, 以防止意外发生并避免不可估量的损失. 现有的研究大多基于重要性采样来解决行为策略和目标策略引起的轨迹分布不匹配问题. 然而, 当问题涉及长视距 (或无限视距) 时, 此类方法可能会面临“视距诅咒”, 即随着时间跨度的延长, 累积重要性比的方差可能会呈指数级增长. 为解决这一问题, Lee 等[37] 提出一种带有平稳分布估计的表征平衡框架 RepB-SDE (Representation balancing with stationary distribution estimation). 该方法首先将状态–动作对输入到前馈网络中, 以获得潜在空间表征. 接着利用潜在动力学模型来预测奖励与下一时刻的状态. 然后, 在表征空间中, 通过正则化数据分布与目标策略引起的折扣平稳分布之间的距离来学习平衡表征. 为训练一个不受视距诅咒影响的环境模型, 进一步提出一种基于状态–动作对的表征平衡目标. 实验结果表明, 使用 RepB-SDE 目标训练的模型对异策略评估任务的分布偏移具有很好的鲁棒性, 尤其是在目标策略和行为策略之间存在较大差异的情况下. 因此, RepB-SDE 为解决强化学习中的视距诅咒问题提供了有效的解决方案.

在监督学习中, 尽管深度神经网络的参数过多, 但其却有很好的泛化性能, 这主要归因于通过随机梯度下降优化器所诱导的隐式正则化的作用, 它使模型倾向于选择具有良好泛化性的简约解. 那么, 这一优势能否推广到深度强化学习领域中呢? Kumar 等[36] 对此进行了探讨, 并给出否定的答案: 监督学习中的隐式正则化效应在离线强化学习环境下却适得其反, 导致较差的泛化与退化的特征表示. 首先, 从理论上证明当现有的隐式正则化模型应用于时序差分学习时, 所得到的正则化器倾向于具有过度“混淆”的退化解, 而不是贝尔曼方程的一个稳定不动点, 这与监督学习的结果形成鲜明对比. 仿真测试也证实了通过自举训练的深度值函数网络学到的特征表示确实会出现退化现象. 为解决该问题, Kumar 等[36] 进一步提出一种简单有效的显式正则化方法, 称为 DR3 (Value-based deep RL requires explicit regularization), 它最小化自举更新中出现的状态–动作对之间的特征相似性, 从而抵消隐式正则化带来的不利影响. 仿真结果表明, 当与现有的离线强化学习方法相结合时, DR3 显著提高了性

能与稳定性. 该文献也给我们提供了一个优化时的技巧, 即可以使用 DR3 正则化器进一步提升离线强化学习算法性能.

实现采样高效的离线策略评估需要满足两个充分条件, 即贝尔曼完备性与数据覆盖性. 先前的离线强化学习方法都是事先假设已给定了满足这两个条件的表征, 其结果基本上都是理论性的. 相比于先前的方法, Chang 等[38] 直接从数据中学习具有良好覆盖性的近似线性贝尔曼完备性表征, 并提出具有贝尔曼完备性和探索性表征学习的离线策略评估 (Offline policy evaluation with bellman complete and exploratory representation learning, BCRL) 算法. 其主要思想为: 利用丰富的函数近似学习贝尔曼完备性与探索性表征, 从而实现稳定和准确的离线策略评估; 使用最小二乘策略评估与所学表征中的线性函数来进行离线策略评估; 对 BCRL 进行端到端的理论分析, 并证明所提评估算法具有多项式数量级的样本复杂度. 在连续控制机器人任务上的仿真结果表明, BCRL 能实现更好的离线策略评估. 同时, 消融实验也说明近似线性贝尔曼完备性与覆盖性是 BCRL 成功的关键因素, 这两个因素缺一不可. 未来可以考虑将 BCRL 扩展到离线策略优化中.

基于状态–动作对表征的离线强化学习能够学习状态与动作的联合表征, 从而捕捉两者之间的交互关系, 并为其后续决策任务提供重要指导. 这类方法已经在基于视觉的机器人控制、自动驾驶[37]、游戏[36]、医疗[63] 等领域得到一定的应用. 然而, 与基于状态表征的离线强化学习类似, 这类方法仍然受限于基于视觉感知的观测输入场景.

2.4　轨迹表征

轨迹表征是指将强化学习问题视作固定经验的条件序列建模问题. 通过使用 Transformer[39−44] 或扩散模型[45−47], 将强化学习中的状态、动作和奖励等数据转化为一系列去结构化的序列数据, 并对这些序列数据进行建模, 以预测未来的轨迹序列, 参考文献 [39], 基于轨迹表征的离线强化学习框架如图 5 所示. 其中, s_t, a_t, r_t 分别表示 t 时刻下的状态、动作与奖励值. 这种方法的优势在于, 它能够避免传统强化学习中的自举误差问题, 同时发挥 Transformer 或扩散模型架构在序列建模和轨迹表征方面的强大能力. 通过轨迹表征, 能够更深入地理解和利用序列数据, 捕捉数据之间的时间关联性, 从而提高算法的泛化性能, 并在控制和决策任务中取得更好的效果.

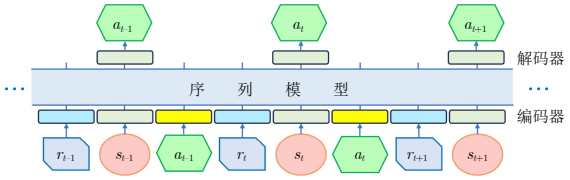


图 5　基于轨迹表征的离线强化学习框架

Fig. 5　The framework of offline reinforcement learning based on trajectory representation

序列模型是指输入或者输出中包含有序列数据的模型, 而深度序列模型则是利用深度神经网络来建模序列的条件分布. 2017 年, 谷歌团队提出一种简单且有效的网络架构 Transformer[64], 该架构完全摒弃了先前神经网络中的递归与卷积操作, 利用注意力机制提高模型的训练速度. 谷歌团队将 Transformer 架构应用于自然语言处理领域的机器翻译任务上, 并取得很好的效果. 由于其具有强大的表征能力、时序建模能力与可扩展性, Transformer 得到了学者们的青睐. 近年来, Transformer 不仅在自然语言处理上成效显著, 而且在强化学习[39−44]、计算机视觉[65] 等领域也取得重大突破. 最近, 有学者开始将其优势与离线强化学习相结合, 在建模序列的条件分布基础上, 根据当前状态采样未来可能的序列, 从而解决离线环境下的决策问题.

不同于上述利用策略约束或值约束设计离线强化学习方法的思路, Chen 等[39] 和 Janner 等[40] 独辟蹊径, 将由状态、动作、奖励和价值构成的轨迹视为一串去结构化的序列数据, 基于 Transformer 模型提出了两种离线强化学习方法: DT (Decision Transformer) 和 TT (Trajectory Transformer).

Chen 等[39] 将离线强化学习视为序列决策问题, 并提出 DT 算法. 该算法不使用传统的强化学习方法 (如时序差分) 求解最优策略, 而是利用序列建模目标在收集而来的离线数据集上训练模型, 并直接进行决策. 通过对期望回报、历史状态与动作序列的自回归模型进行调节, 所提算法能够产生最优动作. 由于 Transformer 强大的表征能力, 其在拟合数据上具有一定的泛化性, 因此在强化学习基准任务上显示 DT 优于现有的无模型离线强化学习算法与行为克隆方法.

同年, Janner 等[40] 提出 TT 算法. 相比于 DT, TT 更侧重于对轨迹分布进行建模. 其主要思路为: 将整个序列中的元素按维度进行离散化处理, 然后使用离散自回归方式来拟合离线数据集中状态、动作与奖励的分布, 并使用束搜索对解码出的候选轨迹进行进一步优化. 在长时域动态预测、模仿学习以及离线强化学习任务中展现出 TT 的灵活性, 同时,

该算法可以与现有的无模型算法相结合, 在稀疏奖励、长时域任务上优于现有的规划方法. 然而, 离散化操作引发了更高的计算复杂度, 因此该模型预测速度慢, 很难应用于实时控制领域.

提高算法的采样效率是强化学习所面临的一大难题. 最近的研究表明, 使用表达性强的策略函数近似器和对未来轨迹信息进行调节能够有效学习多任务策略. Furuta 等[41] 将这些方法归结为事后信息匹配问题, 即使用未来状态信息来自动挖掘最优轨迹数据. 为解决该问题, Furuta 等[41] 在 DT 的基础上做进一步扩展, 提出 GDT (Generalized decision Transformer) 模型. 该模型将原先 DT 所拟合的奖励分布改为轨迹中包含的信息, 以行为克隆为学习目标, 学习一个条件策略来生成满足不同属性 (包括分布) 的轨迹, 从而提高了 Transformer 架构在强化学习中的适用性. 在 MuJoCo 连续控制基准上验证了 GDT 下游算法 (分类 DT 与双向 DT) 的有效性.

基于序列模型的离线强化学习算法, 如 DT、TT 和 GDT, 主要关注如何最大化累积奖励, 而忽略了可能导致智能体失败或产生危险行为的风险[66]. 为确保智能体在学习和决策过程中, 不会产生不安全或不可接受的行为, Liu 等[42] 在 DT 架构的基础上提出约束决策 Transformer (Constrained decision Transformer, CDT) 算法. 该算法能够从离线数据集中学到安全且自适应的策略, 在系统获得高奖励的同时, 也考虑到系统的安全约束和限制, 以减小潜在风险. 其中, 两个关键技术对于提高冲突目标回报的安全性和鲁棒性方面发挥了至关重要的作用. 首先, 基于熵正则化的随机策略方法能够允许策略探索更多样化的动作范围, 并通过与环境的交互提高性能. 其次, 基于目标回报重标记的 Pareto 边界数据增强方法能够解决期望回报之间的潜在冲突, 并确保目标成本优先于目标回报. 实验结果表明, CDT 在学习自适应、安全、稳健和高回报策略方面表现出很大的优势. 更值得一提的是, CDT 无需重新训练策略就可以适应不同的约束阈值, 显示出强大的灵活性和适应性. 然而, CDT 算法缺乏严格的安全理论保障, 并且由于使用了 Transformer 体系结构, 计算资源消耗较大.

上述方法将状态、动作和奖励 3 种模态统一建模为一个序列, 并未考虑序列决策过程中不同模态之间的相互作用. 为此, Wang 等[43] 提出一种多模态序列建模方法 DTd (Decision transducer), 将这 3 种模态的轨迹分解为 3 个单模态轨迹. 通过分析 DT 的最后一层注意力分数 (如回报−状态、状态−动作), DTd 量化了序列决策过程中的跨模态和模态内交互作用. 研究结果表明, 在表征学习过程中, 应首先处理模态之间不太重要的交互, 然后再处理重要的交互, 以确保在表征学习过程中, 将最重要的交互干扰降到最低. 在 D4RL 基准测试中, DTd 不仅优于先前基于 Transformer、时序差分和扩散模型的方法, 而且在训练过程中具有更好的采样效率和算法灵活性.

Transformer 架构的出现引发了强化学习领域的新一轮研究热点, 离线强化学习与 Transformer 相结合的变体更是在很多任务上胜过传统的强化学习算法, 为研究通用决策模型提供了很好的解决思路. 然而, 该架构摆脱了原有强化学习的思维模式, 在大多数情况下, 传统强化学习中的有益成果无法与 Transformer 有效结合, 导致在决策算法设计层面受到很大限制.

为解决这一问题, Zeng 等[44] 进行了初步探索, 研究了序列建模得到的轨迹表征对策略学习的影响, 并提出一种名为目标条件预测编码 (Goal-conditioned predicitve coding, GCPC) 的方法. 该方法将任务划分为两个独立的阶段: 轨迹表征提取与策略学习. 具体而言, 首先, 利用序列建模技术对目标条件下的历史轨迹信息进行压缩和编码, 得到预测状态序列表征. 然后, 将提取的潜在表征传递给监督强化学习模型, 以学习策略与期望目标. 所提出的框架能够灵活地与现有的监督强化学习方法相结合, 使我们能够探索不同的序列决策设计选择对性能的影响. 然而, 这种两阶段方式需要进行特征设计与模型调整, 因此探索自动化水平更高的端到端方式是未来值得研究的课题.

前述 DT 模型是在真实轨迹回报的基础上通过 GPT (Generative pre-trained Transformer) 来学习轨迹生成器. 类似地, Janner 等[45] 和 Ajay 等[46] 利用扩散模型强大的数据生成能力来扩散生成轨迹序列. Janner 等[45] 利用扩散模型作为轨迹生成器对机器人行为进行规划, 并提出一种基于时间卷积的轨迹规划扩散模型, 称为 Diffuser. 该方法将状态−动作对的完整轨迹组合在一起, 将其作为扩散模型的单个样本. 首先学习单独的回报模型来预测每个轨迹样本的累积奖励, 然后在逆向采样阶段引入辅助引导函数, 通过用一些可微目标的梯度扰动网络输出来指导扩散模型的去噪过程, 从而生成优化给定奖励函数的轨迹. 在稀疏奖励与长时域问题中展现出所提框架的有效性. 然而, 在线使用 Diffuser 时, 由于状态是与环境交互得到的, 序列模型无法再从状态自回归中预测动作. 因此, 在评估阶段, 预

测每个状态的完整轨迹时只使用第一个动作, 这会产生很大的计算成本.

离线数据集中存在很多次优轨迹, 这些轨迹可能会对条件扩散模型产生负面影响, 导致其性能下降. 为此, Ajay 等[46] 提出决策扩散器 (Decision Diffuser), 利用奖励标记后的轨迹数据集来扩散学习轨迹的回报条件模型. 具体而言, 该方法在给定当前时刻的状态和模型约束条件的情况下, 使用无分类器引导和低温采样来生成一系列未来状态, 从而指导回报条件模型捕获数据集中的最佳行为, 收集使回报最大化的高似然轨迹. 然后, 将当前与下一时刻的状态输入到逆动力学模型中, 推断得到动作. 实验结果表明, 通过采用无分类器引导与低温采样技术, Decision Diffuser 能筛选出质量较好的状态序列, 从而学到更优的策略. 然而, 由于需要使用神经网络来参数化有条件模型与无条件模型, 因此无分类器引导方式进一步增加了训练成本.

文献 [46] 指出, 在有限数据情况下, 扩散模型很容易出现过拟合问题, 因为训练数据的多样性不足限制了扩散模型的质量, 同时也影响其在新任务上的泛化能力. 为此, Liang 等[47] 提出一种名为自适应扩散器 (AdaptDiffuser) 的扩散进化规划方法. 该方法利用奖励梯度的指导为目标条件任务生成丰富的异构专家数据. 然后, 通过使用判别器筛选出高质量的数据以自我改进扩散模型. 实验结果表明, AdaptDiffuser 能够提高扩散模型对未知离线强化学习决策任务的泛化能力. 然而, 直接从原始高维观测状态中进行扩散学习可能导致训练时间长且效率低下, 从而限制了其在现实领域的应用.

Diffuser、Decision Diffuser 和 AdaptDiffuser 均不是使用传统的强化学习范式 (如时序差分) 来求解最优策略, 而是利用序列建模目标在收集而来的离线数据集上训练模型并直接进行决策. 在实际使用过程中, Diffuser、Decision Diffuser 和 AdaptDiffuser 只需要预测得到的第一个动作或状态即可. 但是, 由于这类轨迹级序列模型无法从状态中自回归地预测动作, 因此有必要将每个状态的整条轨迹预测出来, 这将消耗巨大的额外计算成本. 更为重要的是, Diffuser、Decision Diffuser 和 AdaptDiffuser 相对于传统的强化学习范式来说是比较颠覆的, 基本上完全以深度网络模型为核心. 关于 Diffuser、Decision Diffuser 和 AdaptDiffuser 如何将一个可能不切实际的高目标价值联系到一个合适动作的整个过程完全是黑箱的, 这种黑箱性质直接导致很多已有的强化学习成果难于融入其中.

基于轨迹表征的离线强化学习将强化学习视为条件序列建模问题, 用于预测未来轨迹序列, 在机器人连续控制、游戏[39]、导航[40, 44]、工业芯片布局[67] 等领域已经取得显著成功. 尽管使用扩散模型可以有效地对策略进行建模, 但相比于其他生成模型, 扩散模型的生成速度相对较慢, 因此无法适用于实时控制场景. 如何在不损失性能的情况下提升生成速度是有待解决的问题.

2.5　任务或环境表征

强化学习的核心思想是通过智能体与环境的交互来学习最优策略, 以实现最大化期望回报的目标. 然而, 在实际应用中, 强化学习面临着许多挑战, 其中之一就是如何在不同任务之间实现快速适应. 传统的强化学习方法通常需要针对每个任务单独训练一个模型, 这种方式存在着许多问题. 首先, 不同任务之间的数据往往存在很大差异, 因此需要针对每个任务重新设计模型, 增加了开发成本和难度. 其次, 由于模型只能针对特定任务进行训练, 因此在面对新任务时需要重新训练模型, 这会导致训练时间的延长和性能的下降. 为解决这些问题, 学者们开始探索一种新的强化学习范式, 即离线元强化学习. 离线元强化学习使用元学习的方法, 不仅从历史数据中学习, 还试图发现和学习在各种不同环境下的通用策略. 其目标是使得该策略能够快速适应类似但未被观测到的新任务, 从而提高学习效率和泛化性能. 通过定义编码器, 推断得到任务[48−49] 或环境[50−51] 表征, 训练智能体的目的是根据获取的表征来调整策略的学习, 使智能体能够有效地适应到从未见过的任务中.

参考文献 [68], 基于任务 (环境) 表征的离线强化学习框架如图 6 所示. 其中, c 表示任务 (或环境), 经过编码器, 得到其潜在表征 z. 策略 $\pi_\theta(a|s,\ z)$ 给出了智能体在给定任务 (环境) 表征 z 与观测状态 s 的条件下选择动作 a 的概率, θ 为 Actor 网络参数. 状态动作值函数 $Q_\psi(s,\ a,\ z)$ 用于评估在给定任务 (环境) 表征 z 与状态 s 条件下执行动作 a 的

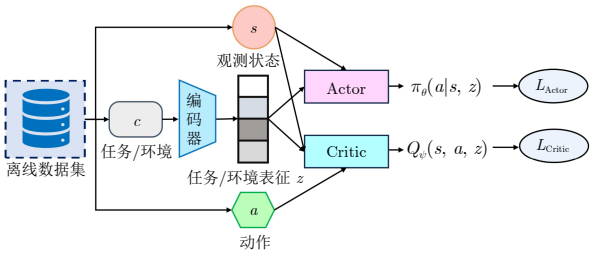


图 6　基于任务 (环境) 表征的离线强化学习框架

Fig. 6　The framework of offline reinforcement learning based on task (environment) representation

价值, ψ 为 Critic 网络参数. L_{Actor} 与 L_{Critic} 分别表示策略网络 Actor 与价值网络 Critic 的损失函数.

离线轨迹数据分布是由行为策略和任务共同决定的. 现有的离线元强化学习算法无法区分这些因素, 使得任务表示对行为策略的变化不稳定. 为此, Yuan 等[48] 研究了从离线多任务数据中学习任务表示与元策略, 并提出用于离线元强化学习的对比稳健任务表示学习 (Contrastive robust task representation learning for offline meta RL, CORRO). 首先, 将算法框架建模为一个双层结构的任务编码器, 其中第 1 层是从一步转移元组而不是整个轨迹中提取任务表示, 第 2 层则是对这些表示进行聚合; 然后, 将表示与任务之间的最大化互信息作为学习目标, 从而最大限度地消除任务表示中行为策略的影响; 最后, 引入对比学习来优化互信息下界, 并给出两种近似负样本对真实分布的方法, 包括生成建模与奖励随机化方法. 在 Point-Robot 与多任务 MuJoCo 基准测试下的仿真结果表明, 相较于先前基于上下文的离线元强化学习方法, CORRO 能够更好地泛化到分布外的行为策略上.

上述 CORRO[48] 使用监督对比学习训练任务编码器, 将相同任务样本视为正样本对, 而所有其他样本视为负样本对, 通过对正样本进行聚类并在嵌入空间中排除负样本来学习数据表征. 然而, 这种方法容易受到元训练和元测试阶段使用的行为策略之间可能产生的分布差异的影响. 实验结果表明[49], CORRO[48] 无法很好地适应低质量任务, 当使用与元训练策略不同的策略预先收集任务数据信息时, CORRO 的性能会急剧下降. 这是由于训练的离线数据主要由高质量专家数据组成, 导致 CORRO 无法捕获到低质量任务数据中包含的信息. 为解决这个问题, Zhao 等[49] 提出一种基于硬采样 (Hard sampling) 的策略来训练鲁棒的任务编码器, 称之为 HS-OMRL. 该方法采用以任务索引为标签指导的监督式对比学习框架. 假设在元测试阶段会遇到各种质量的任务, 根据嵌入空间样本之间的距离来衡量正负样本的"硬度"值, 从而调整监督对比损失函数. 实验结果表明, 与现有方法相比, HS-OMRL 可以获得更鲁棒的任务表征.

目前大部分离线强化学习方法都是基于保守主义的思想, 将学习严格限制在离线数据集的支持域内, 从而避免产生分布外动作. 但这种做法只会获得局部次优策略. Chen 等[50] 直接研究了支持域外的决策行为, 提出一种新的离线强化学习框架, 称为基于模型的自适应策略学习 (Model-based adaptable policy learning, MAPLE). 其基本思路为: 利用集成技术构造所有可能的动力学模型, 从而建模分布外的区域; 基于元学习的思想, 引入额外的环境特征提取器来提取环境表征, 所学策略根据该环境表征进行自适应调整. 通过这种方式, MAPLE 能够学到一个自适应策略, 该策略在部署时可以适应支持域外的动作. 在离线控制任务中进行仿真, 结果表明, MAPLE 能在支持域外做出稳健的决策, 与现有的离线强化学习算法相比, 所提算法在大多数任务上都获得了最优的性能. 该算法也给予我们一些启发: 通过对动力学模型进行扩充, 并使用自适应策略对模型的探索边界进行拓宽, 从而获得更稳健的决策. 然而, 环境特征提取器的泛化能力依赖于神经网络本身, 能否提升其在未见模型中的泛化性能, 是未来值得研究的方向.

为解决只有离线训练环境经验数据情况下智能体的快速适应问题, Sang 等[51] 提出一种用于快速策略适应的离线强化学习方法 PAnDR (Policy adaptation with decoupled representations). 该方法分为离线训练和在线适应两个阶段. 在离线训练阶段, 智能体从不同环境中收集到的离线经验数据中进行学习. 首先, 使用轨迹级对比学习和策略恢复分布学习环境表征和策略表征; 然后, 利用互信息进一步优化表征, 减少环境表征和策略表征之间的冗余信息; 接下来, 基于学到的表征, 训练策略–动力学值函数网络来近似不同策略和环境组合的值. 在在线适应阶段, 根据从新环境中收集的少量经验数据推断出环境上下文, 并利用梯度上升进行策略优化. 实验结果表明, PAnDR 在几个代表性的策略适应问题上展现出优越性. 然而, 由于缺乏对空间的全局了解, 优化策略表征并非易事. 同时, 如果不仔细处理不确定性, 对得到的潜在空间表征进行解码可能是无效的.

基于任务或环境表征的离线强化学习借助元学习的思想, 使其所学习到的策略能够在新任务上展现出良好的泛化性能. 目前, 基于任务或环境表征的离线强化学习已经在机器人连续控制、导航[48−49]、智能驾驶[69]、量化交易[70] 等领域取得了一定的突破. 然而, 策略的泛化能力严重依赖于任务 (环境) 之间的相似性. 如果新任务 (环境) 与已有任务 (环境) 存在较大差异, 则可能无法有效地迁移知识和经验. 因此, 需要进一步研究如何提高策略的泛化能力, 以适应更加复杂的任务 (环境).

3 实验设置

近年来, 离线强化学习算法已取得一定进展, 因此需要有统一的离线数据集与评价指标来对不同

算法进行比较, 以此来衡量算法的真实性能. 为此, 本节介绍 3 种常用的离线强化学习基准数据集 RL Unplugged[71]、D4RL[72] 和 NeoRL[73], 并对现有的离线策略评估与超参数选择方法进行综述, 离线强化学习基准数据集对比如表 2 所示.

3.1　基准数据集

3.1.1　RL Unplugged

Gulcehre 等[71] 提出一个名为 RL Unplugged 的基准套件来评估与比较离线强化学习算法性能. RL Unplugged 包括来自不同领域且形式多样的数据集、详细的评估协议以及现有几种离线强化学习算法的评估结果. 通过使用基准数据库, 可以提高实验的复现性、算法对比的系统化、规范化, 使学者们能在有限的计算成本下研究具有挑战性的任务. RL Unplugged 开源代码库: https://github.com/deepmind/deepmind-research/tree/master/rl_unplugged.

1) 领域、任务与数据集

RL Unplugged 包括来自不同领域的数据, 涵盖游戏与模拟运动控制问题, 数据集形式多样, 包含部分与完全可观测、离线与连续动作、具有随机性与确定性动力学的任务.

a) DM Control Suite

连续动作域, 包含 9 个任务, 其中 5 个用于在线策略选择, 4 个用于离线策略选择. 数据集生成方式: 首先, 大部分数据集由 D4PG 生成, 对于 Manipulator insert ball 与 Manipulator insert peg 这两个任务, 由 V-MPO[74] 生成; 其次, 在所有任务上每种算法均独立重复 3 次来确保数据的多样性, 且记录整个训练过程中的数据; 最后, 由于离线方法通常处理数据量少的情况, 因此通过下采样来减少数据量, 同时, 将每个数据集中成功轨迹的数量减少 2/3, 以确保数据集不包含太多的成功轨迹.

b) DM Locomotion

连续动作域, 包含 7 个任务, 其中 4 个用于在线策略选择, 3 个用于离线策略选择. 数据集生成方式为: 首先, 对于 3 个 Humanoid 任务, 使用文献 [75] 训练的专家策略, 每个任务只有一个动作捕捉的运动技能模块被重复使用, 且数据集由 3 种在线方法生成. 对于 4 个 Rodent 任务, 使用与文献 [76] 一致的训练方法, 且数据集由 5 种在线方法生成. 其次, 记录整个训练过程中的数据, 对其进行下采样, 且成功轨迹减少 2/3. 最后, 由于环境感知是由以自我为中心的摄像机完成的, 因此运动任务中的所有数据都包含每个时间间隔上大小为 $64 \times 64 \times 3$ 的摄像机观测画面, 使用以自我为中心的观测可能会使一些环境具有部分可观测性, 且需要循环架构, 为此生成了序列数据集.

c) Atari 2600

离散动作域, 包含 46 个任务, 其中 9 个用于在线策略选择, 37 个用于离线策略选择. 任务选取的

表 2　离线强化学习基准数据集对比

Table 2　Comparison of benchmarking datasets for offline reinforcement learning

名称	领域	应用领域	数据集特性
RL Unplugged	DeepMind 控制套件	机器人连续控制	连续域, 探索难度由易到难
	DeepMind 运动套件	模拟啮齿动物的运动	连续域, 探索难度大
	Atari 2600	视频游戏	离散域, 探索难度适中
	真实世界强化学习套件	机器人连续控制	连续域, 探索难度由易到难
D4RL	Maze2D	导航	非马尔科夫策略, 不定向与多任务数据
	MiniGrid-FourRooms	导航, Maze2D 的离散模拟	非马尔科夫策略, 不定向与多任务数据
	AntMaze	导航	非马尔科夫策略, 稀疏奖励, 不定向与多任务数据
	Gym-MuJoCo	机器人连续控制	次优数据, 狭窄数据分布
	Adroit	机器人操作	非表示性策略, 狭窄数据分布, 稀疏奖励, 现实领域
	Flow	交通流量控制管理	非表示性策略, 现实领域
	FrankaKitchen	厨房机器人操作	不定向与多任务数据, 现实领域
	CARLA	自动驾驶车道跟踪与导航	部分可观测性, 非表示性策略, 不定向与多任务数据, 现实领域
NeoRL	Gym-MuJoCo	机器人连续控制	保守且数据量有限
	工业基准	工业控制任务	高维连续状态和动作空间, 高随机性
	FinRL	股票交易市场	高维连续状态和动作空间, 高随机性
	CityLearn	不同类型建筑的储能控制	高维连续状态和动作空间, 高随机性
	SalesPromotion	商品促销	由人工操作员与真实用户提供的数据

原则为: OpenAI Gym 中涵盖的 Atari 游戏超过 46 个, 这里仅挑选出使用在线 DQN 方法比使用随机策略方法性能好的那些游戏环境; 根据游戏难度对所有游戏进行排序, 挑选每 5 个游戏作为离线策略部分的任务, 这样就能覆盖到不同难度的游戏集合. 数据集生成方式: 运行一个在线 DQN 智能体并在训练期间使用粘性动作从经验回放池中记录转移元组[77]; 每个游戏都运行 5 次, 每次含有 5000 万次转移元组, 每个转移元组中的状态都包含 4 个帧的堆栈, 以便能够与基准进行帧堆栈.

d) Real-world RL Suite

连续动作域, 包含 4 个任务. 数据集生成方式为: 按照文献 [78] 的方式, 运行无挑战设置与组合挑战设置下 (除安全与多目标奖励外) 的次优策略, 从而得到离线数据. 其中次优策略的获得方式为: 在分布最大化后验策略优化算法[79] 上采用不同的随机权重初始化来训练 3 个智能体直到收敛, 然后根据大约 75% 的收敛性能进行拍快照从而得到策略. 对于无挑战设置, 组合 3 个快照为每个环境生成 3 个大小不同的数据集. 组合挑战设置与此类似, 为确保任务仍然能够得到解决, 对组合挑战使用 “large data”.

2) 基准算法

包括行为克隆 (BC[80])、在线强化学习算法 (DQN[81]、D4PG[82]、IQN[83]) 与离线强化学习算法 (BCQ[15]、BRAC[84]、ABM[85]、REM[86]).

3) 评估协议

在线评估协议. 能够以在线的方式与环境互动, 通过环境反馈并计算得到的期望回报来评估不同超参数下算法的性能. 但这种方式在许多实际领域中是不可行的.

离线评估协议. 完全离线, 不与环境进行交互, 仅依靠离线数据来评估不同超参数设置下策略的好坏. 由于无法像在线策略选择那样, 可以在环境中运行不同超参数所获得的策略, 并通过回报来判断该选择何种策略, 因此这种方式具有一定难度.

仿真测试结果表明, 离线强化学习在一些控制套件任务与 Atari 游戏中表现很好, 但在部分可观测环境中, 如运动套件, 离线强化学习方法的性能较差. 因此, 如何缩小现实世界与模拟环境之间的差距, 仍是离线强化学习未来具有挑战性的问题.

3.1.2 D4RL

以离线强化学习实际应用相关数据集的关键属性信息为指导, Fu 等[72] 引入专门为离线环境设计的基准数据集 D4RL. 与 RL Unplugged 不同的是, Fu 等[72] 将设计重点放在任务与数据收集策略上, 这些任务与数据收集策略涵盖了实际离线场景中可能面临的挑战. 为便于学者们进一步研究, D4RL 中涉及到的基准任务、数据集、现有算法、评估协议等均已开源, D4RL 开源代码库: https://github.com/rail-berkeley/d4rl.

1) 离线数据集设计准则

为模拟真实世界的离线数据集, Fu 等[72] 指出, 所构造的离线数据集应涵盖如下几种类型.

a) 狭窄和有偏数据分布. 离线强化学习一个重要挑战是能够处理不同的数据分布, 而算法不会偏离或产生比提供的行为更差的性能. 处理此类数据分布一般采用保守方法, 试图使行为接近数据分布.

b) 不定向与多任务数据. 离线数据记录可能并不全是一条条完整的轨迹, 而是由大量轨迹片段组成, 单个子轨迹可能无法完成一项完整任务, 但可以通过对多个轨迹片段拼接来完成, 因此这些子轨迹仍能为我们提供有用的信息.

c) 稀疏奖励. 由于信用分配问题与探索的困难性, 稀疏奖励问题对在线强化学习方法构成挑战. 在离线强化学习中, 解决稀疏奖励问题仅需考虑算法执行信用分配的能力即可.

d) 次优数据. 对于目标明确的任务, 数据集可能并未包含最优轨迹, 这对模仿学习等需要专家示范的方法来说具有一定挑战.

e) 非表示性行为策略、非马尔科夫行为策略以及部分可观测性. 现实生活中的行为可能并非来自模型类中的策略, 如由人类示范或手工设计的控制器所产生的数据可能落在模型类外, 这样就会引入额外的表示错误. 更为普遍的是, 在假设数据由马尔科夫策略生成的情况下, 当估计具有部分可观测性的非马尔科夫策略与任务的动作概率时, 就会引入额外的建模误差. 这些误差会给离线强化学习算法造成额外的偏差, 尤其是假设从马尔科夫策略中获得动作概率的方法, 如重要性加权.

f) 现实领域. 在真实世界环境下进行评估是对离线强化学习进行基准测试的理想设置, 但这与一个可广泛获取且可复现的基准不一致. 为达到平衡, D4RL 选择在在线强化学习中广泛使用的模拟环境 (如 MuJoCo、Flow、CARLA 等). 此外, 在某些领域, 还利用了人类示范或人类行为的数学模型, 从而提供从现实过程中产生的数据集.

g) 多种类型且难易不同的任务. D4RL 包含各种类型的任务: 机器人操纵、控制、导航、运动、交通管理、自动驾驶等. 此外, 每项任务中都涵盖不同的难易程度, 从当前算法能够解决的任务到目前无法解决的难度系数更高的问题.

2) 领域、任务与数据集

基于上述原则, D4RL 提供来自 8 个领域的 58 项任务. 其中, Maze2D 与 MiniGrid-FourRooms (Maze2D 的离散模拟) 为导航任务, 由非马尔科夫策略、不定向与多任务数据构成. AntMaze 为导航任务, 由非马尔科夫策略、稀疏奖励以及不定向与多任务数据构成. Gym-MuJoCo 为连续控制任务, 由次优数据与狭窄数据分布构成. Adroit 为机器人操作任务, 由非表示性策略、狭窄数据分布、稀疏奖励以及来自现实领域的数据构成. Flow 为交通流量控制管理任务, 由非表示性策略与来自现实领域的数据构成. Franka-Kitchen 为厨房机器人操作任务, 由不定向与多任务数据以及来自现实领域的数据构成. CARLA 为自动驾驶车道跟踪与导航任务, 由部分可观测性、非表示性策略、不定向与多任务数据以及来自现实领域的数据构成.

3) 基准算法

包括 BC[80], 在线与离线 SAC 算法 (SAC 与 SAC-off)[87], 策略约束方法 BCQ[15]、BEAR[88]、BRAC-p[84]、BRAC-v[84]、AWR[89], 值函数正则化方法 CQL[90], 边际重要性采样方法 AlgaeDICE[91], 不确定性估计方法 Continuous REM[86].

4) 评估协议

D4RL 采用离线评估方式对算法性能进行评估. 在每个领域中指定一个任务子集作为训练任务, 允许其进行超参数调优, 另一个子集作为评估任务, 用于衡量最终性能. 为便于在任务之间进行比较, 计算归一化分数 (NS), 计算式为

$$NS = 100 \times \frac{score - random\ score}{expert\ score - random\ score} \tag{4}$$

其中, $score$ 为当前分数, $random\ score$ 为随机分数, $expert\ score$ 为专家分数. 归一化分数为 0 表示在动作空间中智能体完全随机采取动作得到的平均回报 (超过 100 回合), 分数为 100 表示完全来自特定领域专家得到的平均回报.

仿真测试结果表明, 很多算法在控制器生成数据的任务上取得一定程度的成功, 如 Flow 与 CARLA. 但在数据有限的任务上, 如 Adroit 和 Franka-Kitchen 的人类示范数据, 现有的离线强化学习算法仍具有挑战性. 由于在很多实际情况下大规模数据集并不总是可获取的, 因此未来研究更高效的采样方法仍是相当必要的.

3.1.3 NeoRL

RL Unplugged 仅使用在线训练的经验回放池来构造数据集, 相比之下, D4RL 提供了多种类型的数据集, 包括随机生成、带有经验回放池、专家策略、混合策略、人类示范等收集而来的数据. 尽管 RL Unplugged 与 D4RL 在实现测试与比较离线强化学习算法方面做出了巨大贡献, 但在更一般化的现实应用领域上 (如推荐系统、产品促销、工业控制等), 先前的基准数据集仍面临一定挑战: 首先, 在很多实际情况下, 为确保系统的安全性, 无法通过运行过度探索性策略来收集大量数据, 因此只能使用狭窄的数据分布; 其次, 只有当所学策略优于实际部署在系统中的策略时, 才认为此策略有效; 最后, 对所学策略进行离线策略评估是非常必要的, 只有得到充分验证后, 所学策略才能部署到真实环境中. 为应对上述挑战, Qin 等[73] 提出一个接近真实世界的离线强化学习基准 NeoRL. 相比于其他基准数据集, NeoRL 提供了更具挑战性和多样化的任务, 以更好地模拟真实世界的复杂性和随机性. NeoRL 数据集是通过收集更加保守的策略而形成的, 涵盖机器人控制、工业控制、市场营销等高维度或者具有高度随机性的真实场景任务, 这些任务旨在帮助研究人员和开发者在复杂的现实环境中测试和评估离线强化学习算法的性能. 相关基准任务、数据集、现有算法、评估协议等可见 NeoRL 开源代码库: https://github.com/polixir/NeoRL.

1) 离线数据集设计准则

首先, 获取数据收集策略. 使用 SAC[87] 对所有环境进行训练直至收敛, 且每一轮均记录一个策略. 将整个训练过程中具有最高回报的策略作为专家策略, 进一步对 3 个等级的策略 (约为 25%、50%、75% 专家性能的策略) 进行存储, 从而模拟多级次优策略, 分别用低 (L)、中 (M)、高 (H) 表示.

其次, 收集数据. 当概率为 20% 时, 从训练好的高斯策略中采样, 否则, 使用高斯均值策略.

最后, 划分训练与测试数据集. 对于每个等级, 选择具有相似回报的 4 个策略, 其中随机选择 3 个策略来收集用于离线强化学习策略训练的训练数据, 剩余的 1 个策略生成测试数据用于离线验证.

这里测试数据集的默认大小为每个任务中训练数据的 1/10. 额外的测试数据集可用来设计离线评估方法, 如训练过程中的模型与超参数选择. 此外, 对于每个任务, 默认情况下, 提供 3 种尺寸的训练数据集, 分别为 10^2 、10^3 和 10^4 的轨迹量.

2) 领域、任务与数据集

该基准数据集涵盖了 5 大领域, 共计 52 项任务. 其中, Gym-MuJoCo 环境包含了 3 个机器人连续控制任务, 这些任务的数据集是通过采用保守的策略收集而来的, 并且数据量有限. 而 IB、Fin-

RL 与 C-ityLearn 三个领域的数据集均来自高维连续状态和动作空间, 并且具有高度的随机性. 其中, IB 领域用于模拟各种工业控制任务的特性, 例如风力、燃气轮机或化学反应器; FinRL 模拟股票交易市场, 其涵盖了过去 10 年中 30 只股票的交易历史数据; C-ityLearn 则用于控制不同类型建筑的储能, 以重塑电力需求的聚集曲线. 除此之外, SalesPromotion 模拟真实的商品促销平台, 其目标是实现平台运营商总收入的最大化, 其数据集的提供者为人工操作员和真实用户.

3) 基准算法

包括行为克隆 BC、无模型离线强化学习算法 BCQ[15]、PLAS[20]、CQL[90] 与 CRR[92], 以及基于模型的离线强化学习算法 BREMEN[93] 与 MOPO[94].

4) 评估协议

NeoRL 首先选用两种有代表性的离线策略评估方法 FQE (Fitted Q evaluation)[95] 与 WIS (Weighted importance sampling)[96]. 其中, FQE 以策略为输入, 通过贝尔曼回溯对固定数据集进行策略评估. 学习策略的 Q 函数后, 其性能由数据集的初始状态和策略动作的平均 Q 值来衡量. WIS 是对重要性采样 (Importance sampling, IS) 的进一步改进. IS 仅使用目标策略和行为策略之间的比例来对回报进行加权, 但重要性权重的连乘会引发高方差问题, 因此 WIS 通过简化公式, 减小重要性权重, 从而降低 IS 方差.

除根据离线策略评估直接选择最优策略外, 还考虑下面两种指标来评估离线强化学习方法.

Spearman 秩相关系数. 该指标衡量估计值对策略进行排序的程度, 其定义为离线策略评估估计值与真实值有序排序之间的相关系数. 若秩是均匀随机的, 则得分为 0.

Top-K 分数. 该指标衡量离线策略评估选择 K 个策略的相对性能. 首先, 计算所有算法在整个候选策略集中的最大值与最小值, 并将每个策略的实际在线性能归一化为 [0, 1]. 然后, 令 π^k_{off} 为离线策略评估排序得到的第 k 个策略, 则 $\frac{1}{K}\sum_{k=1}^{K}\pi^k_{\text{off}}$ 与 $\max_k\{\pi^k_{\text{off}}\}$ 分别表示为平均分数与最大化前 K 个策略分数, 这里取 $K\in\{1, 3, 5\}$.

仿真测试结果表明, 相比于行为克隆与确定性策略, 部分离线强化学习算法的性能较差, 这意味着如果将这些算法部署到实际应用中, 其性能可能会低于先前的基准测试结果. 而且, 现有的离线策略评估方法很难选出最优策略. 因此, 设计更好的离线策略评估方法仍是未来研究的方向.

3.2 离线策略评估与超参数选择

3.2.1 离线策略评估

离线强化学习面临一大挑战是离线策略评估 (Off-policy evaluation, OPE) 问题, 即仅利用离线数据来评估策略的期望性能[97]. 若能解决此问题, 则在实际部署离线强化学习算法之前就能为使用者提供高可信度保证, 且实现策略改进与超参数选择. 为此, Fu 等[98] 提出深度离线策略评估 (Deep off-policy evaluation, DOPE) 基准, 采用以下 3 个评价指标来衡量策略性能:

1) 绝对误差 ($AbsErr$). 衡量估计精度, 使用绝对误差代替均方误差, 其优势在于对异常值不敏感. 其定义为策略真实值 V^π 与估计值 $\hat{V}^\pi$ 之差的绝对值, 即

$$AbsErr = \left|V^\pi - \hat{V}^\pi\right| \tag{5}$$

2) $Regret@k$. 衡量估计得到的最优策略与整个集合中的最优策略的差值. 它是根据估计的回报识别前 k 个策略来计算的. $Regret@k$ 是整个集合中最优策略的真实期望回报与前 k 个集合中最优策略的真实值之差, 定义为

$$Regret@k = \max_{i\in 1:N} V_i^\pi - \max_{j\in topk(1:N)} V_j^\pi \tag{6}$$

其中, $topk(1:N)$ 表示由估计值 $\hat{V}^\pi$ 衡量得到的前 k 个策略, N 为集合中估计值的个数.

3) Spearman 秩相关系数 ($RankCorr$). 该指标衡量的是估计值对策略进行排序的程度, 其定义为 OPE 估计值与真实值有序排序之间的相关系数, 计算式为

$$RankCorr = \frac{Cov(V^\pi_{1:N}, \hat{V}^\pi_{1:N})}{\sigma(V^\pi_{1:N})\sigma(\hat{V}^\pi_{1:N})} \tag{7}$$

其中, $Cov(\cdot)$ 表示协方差, $\sigma(\cdot)$ 为标准差.

仿真结果表明, 没有任何算法能在所有任务中都表现良好, 且在某一指标 (如绝对误差) 上表现不佳的算法可能在其他指标 (如等级相关系数) 上表现良好. 这些观测结果都可以为从业人员提供参考, 辅助他们根据实际情况选择合适的算法.

由于无法与环境交互, 因此数据集成为离线强化学习算法唯一的信息源, 其决定了所学策略的性能. 只有了解数据集的特性, 以及它如何影响离线强化学习, 才能“对症下药”, 进一步改善算法性能. 为此, 针对离散动作环境, Schweighofer 等[99] 将数据集对离线强化学习算法性能的影响进行全面实证分析. 首先通过 5 种不同的方式生成离线数据集: 随机生成、专家策略、将随机与专家数据集进行混

合、在专家数据集上添加噪声以及在线训练下的回放数据集. 其次, 数据集的特性取决于环境和生成数据集的策略. 为描述不同环境和生成策略的数据集的特性, 使用两个指标来评估数据集:

1) 轨迹质量 (Trajectory quality, TQ). 通过计算轨迹的回报并将其与最大可实现回报的轨迹进行比较来衡量平均数据集回报率.

2) 状态–动作覆盖度 (State-action coverage, SACo). 通过计算数据集中访问的唯一状态–动作对的数量相对于可能的状态–动作对总数来衡量状态–动作空间的覆盖性.

若一个数据集的轨迹平均获得了高回报, 则它就具有较高的 TQ; 若数据集的轨迹覆盖了状态–动作空间中的绝大部分, 则它就具有较高的 SACo. 最后, 在 6 个不同环境中对几种离线强化学习算法进行测试, 结果表明, 异策略深度 Q 网络的离线版本需要具有高 SACo 的数据集才能获得良好的性能, 将目标策略约束到给定数据集这类的算法在具有高 TQ 或 SACo 的数据集上表现良好. 而与离线强化学习算法相比, 行为克隆在具有高 TQ 的数据集上能获得最优或同等的性能. 文献 [99] 对各种离线强化学习算法进行了全面的研究, 以便了解数据集特性对其性能的影响, 但该研究局限于离散状态空间, 在连续状态空间是否具有同样的结果有待进一步探索.

在离线强化学习过程结束后, 需要对不同超参数下获得的策略进行评估, 选出最优策略. 因此针对策略选择问题, 需要设计相应的离线策略评估方法. Konyushkova 等[100] 提出主动离线策略选择 (Active-offline policy selection, A-ops) 算法, 将离线数据与在线交互相结合, 从而确定最优策略. 首先, 提出一种具有扩展观测模型的贝叶斯优化求解方案, 将离线与在线策略评估相结合; 其次, 设计一种高斯过程核来对策略的相关性进行建模, 通过策略所采取的动作来捕获策略之间的依赖关系. 具体而言, 学习高斯过程代理函数, 该函数将策略映射到期望回报; 在高斯过程统计量的基础上构建一个采集函数, 以决定下一步要测试的策略. 在 3 个控制领域的多个任务上进行测试, 结果表明, 与现有的离线与纯在线策略评估方法相比, 所提算法在策略选择质量上有所提升, 但所需的计算成本相对较高.

尽管很多研究致力于评估离线强化学习方法, 但由于所解决的问题不同, 亦或环境与背景不同, 在运用这些在线评估方法时可能会产生不同的最优策略. 如果对在线评估的策略数量不加以限制, 则会与离线强化学习希望减少和环境交互的目标相违背. 为此, Kurenkov 等[101] 提出一种新的评估范式, 称之为期望在线性能 (Expected online performance, EOP). EOP 引入了在线预算的概念, 即在线部署的策略数量, 使用有限的预算对训练好的策略进行离线选择与在线评估, 从而在不同的在线预算约束下找到一组使性能最佳的超参数设置. 从仿真结果中还能得出: 1) 在有限的评估预算下, 行为克隆往往性能最佳; 2) 在线策略选择方法的偏好也与预算有关. 尽管 EOP 强调了在线评估的重要性, 但该方法并未对安全性与风险性进行评估, 在风险敏感的场景下, EOP 的适用性则会受限.

3.2.2 超参数选择

近些年, 基于模型的离线强化学习已取得一定进展. 与无模型的方法不同, 基于模型的离线强化学习从离线数据中训练出动力学模型, 并利用该模型来进一步优化策略. 但现有的方法理论与实践并不完全一致, 理论上悲观回报应受限于模型与真实动力学之间的总变差距离, 但实际却使用模型的不确定性惩罚奖励来实现, 这就催生了各种不确定性启发式算法, 但不同方法之间几乎无法比较. Lu 等[102] 对这些启发式算法进行对比, 利用贝叶斯优化算法来评估不同超参数的选取对性能的影响, 如模型数量或模型产生的轨迹时域等. 通过仿真实验, 得到如下结论: 具有更大惩罚的更长时域的轨迹能够提升现有方法性能; 使用不确定性估计规范形式的惩罚与分布外动作度量实现了更好的相关性; 不确定性与动力学误差的相关性大于分布偏移. 最后, 对关键超参数微调, 使现有的基于模型的离线强化学习方法在大多数基准测试中都获得了统计意义上的性能提升. 这些研究结果可为研究基于模型的离线强化学习算法的从业人员提供实际指导.

折扣因子在提高在线强化学习的采样效率与估计精度方面发挥了重要作用, 但在离线强化学习中的作用尚未得到充分研究. Hu 等[103] 通过理论分析了折扣因子在离线强化学习中的两种不同效应, 即正则化效应与悲观效应. 一方面, 折扣因子是一个调节参数, 用来权衡最优性与采样效率. 另一方面, 较小的折扣因子类似于基于模型的悲观主义, 相当于在最差的模型中使值函数最大化. 进一步, 在线性马尔科夫决策过程背景下对上述两种效应进行定量分析, 并得出两种不同的性能界限. 最后, 在表格马尔科夫决策过程与 D4RL 基准上验证了这两种效应. 仿真结果还表明, 在离线强化学习实际应用中, 较低的折扣因子能带来性能的提升. 但这一现象能否有更好的理论解释还有待进一步研究.

4 应用

基于表征学习的离线强化学习算法应用广泛, 除上述常见的机器人连续控制任务[71−73] 与视频游戏[39, 42] 外, 在工业、推荐系统、智能驾驶等领域也有所涉及, 各类方法的特点如表 3 所示.

4.1 工业

在工业制造领域, 零部件配对与连接器插入是极为常见的. 工业连接器一般是指用于连接工业设备和机器的机电元件, 能够适应恶劣的工业环境并保证电气信号或电力信号的传输可靠性. 工业连接器插入任务指的是在确保工业插头与插座孔位相匹配的前提下, 将连接器插入对应的连接插座中, 以建立电气或信号连接. 随着人工智能应用的蓬勃发展, 机器人硬件成本也随之减低, 因此使用工业机器人替代人工来操作这些任务越来越成为新趋势.

现有的强化学习方法可以解决一部分操作任务, 但通常需要大量实验与漫长的探索过程. 为使工业机器人快速适应不同的工业插入任务, Zhao 等[68] 提出具有示范适应性的离线元强化学习 (Offline meta RL with demonstration adaptation, ODA) 算法. 该算法将上下文元学习与直接在线微调相结合, 从离线数据中元学习自适应策略, 并根据用户提供的少量示范数据来快速适应新任务. 若新任务与先前离线数据集中的任务相似, 则上下文元学习器会立即适应; 若差距较大, 则对策略进行在线微调使其逐渐适应. 使用 KUKA iiwa7 机器人在 9 个工业插入任务中进行测试, 结果表明, ODA 能快速适应各种不同的工业连接器插入任务, 仅使用从头开始学习任务所需的一小部分样本, 就能达到 100% 的成功率.

同样为提升策略的泛化能力, 使机器人能够快速适应未曾见过的工业连接器插入任务, Nair 等[104] 从视觉感知的角度出发, 将域对抗神经网络的域不变性和变分信息瓶颈的域特定信息流控制相结合, 提出一种用于实现奖励模型和策略泛化的表征学习方法 DAIB (Domain adversarial information bottleneck). 该方法的目标是学习一个鲁棒的奖励函数, 用于检测机器人在未见过的连接器上是否成功插入. 然后, 对一种适合在线微调的离线强化学习算法 IQL (Implicit Q-learning)[105] 进行修改, 使其能够与 DAIB 相结合使用. 通过在 50 个不同的连接器上对两个自由度为 7 的 Sawyer 机器人进行预训练, 利用学到的奖励函数可以成功地微调到新的连接器上. 这种方法使机器人能够快速适应新的工业任务, 并具有更好的泛化能力.

芯片布局是芯片设计过程中的关键步骤, 它决定了芯片上各组件之间的相对位置和连线方式. 研究表明, 强化学习可以改善芯片布局的性能. 然而, 传统的在线强化学习需要与环境进行昂贵且耗时的在线交互来从头开始学习, 训练时间长且效率低下. 为此, Lai 等[67] 提出一种名为 ChiPFormer 的离线强化学习算法, 将芯片布局视为序列决策问题. 具体而言, ChiPFormer 使用 GPT 架构作为主干, 该架构采用因果自注意力掩码, 并通过自回归输入标记 (包括电路标记、状态标记和动作标记) 来预测动作. 这种方法能够以数据驱动的方式从固定的离线数据中学习可迁移的布局策略, 并且可以将学到的策略迁移到未见过的芯片电路任务中. 在实际工业

表 3 基于表征学习的离线强化学习应用综述

Table 3 Summarization of the applications for offline reinforcement learning based on representation learning

应用领域	文献	表征对象	表征网络架构	环境建模方式	所解决的实际问题	策略学习方法
工业	[68]	任务表征	编码器架构	无模型	工业连接器插入	从离线数据中元学习自适应策略
	[104]	任务表征	编码器架构	无模型	工业连接器插入	利用域对抗神经网络的域不变性和变分信息瓶颈的域特定信息流控制来实现策略泛化
	[67]	轨迹表征	Transformer	序列模型	工业芯片布局	采用因果自注意力掩码并通过自回归输入标记来预测动作
推荐系统	[57]	动作表征	VAE	基于模型	快速适应冷启动用户	利用逆强化学习从少量交互中恢复出用户策略与奖励
	[60]	状态表征	编码器架构	基于模型	数据稀疏性	利用群体偏好注入的因果用户模型训练策略
	[61]	状态表征	编码器架构	无模型	离线交互推荐	利用保守的 Q 函数来估计策略
智能驾驶	[58]	动作表征	VAE	无模型	交叉口生态驾驶控制	利用 VAE 生成动作
	[69]	环境表征	VAE	基于模型	长视域任务	利用 VAE 生成动作
医疗	[63]	状态–动作对表征	编码器架构	基于模型	个性化诊断	使用在线模型预测控制方法选择策略
能源管理	[59]	动作表征	VAE	无模型	油电混动汽车能源利用效率	利用 VAE 生成动作
量化交易	[70]	环境表征	编码器架构	无模型	最优交易执行的过拟合问题	利用时序差分误差或策略梯度法来学习策略

芯片任务中, ChiPFormer 能够实现更好的布局质量, 并将运行时间缩短了 10 倍.

4.2 推荐系统

在传统的基于强化学习的推荐方法中, 通过每个用户与推荐系统之间产生的交互信息来学习一种能够推断出用户偏好的推荐策略, 从而为不同的用户提供个性化服务. 为学到稳健的推荐策略, 需要用户频繁与推荐系统交互来收集大量数据. 但对于冷启动用户而言, 他们与系统的交互次数有限, 在数据不充足的情况下, 传统的基于强化学习的推荐方法很难从有限的用户–项目交互中挖掘出用户偏好, 从而无法实现个性化的推荐策略. 为解决个性化推荐系统中的用户冷启动问题, Wang 等[57] 利用元学习的思想, 提出基于互信息正则化元模型的强化学习方法 M^{3}Rec (Mutual information regularized meta-level model-based RL approach for cold-start recommendation). 其主要思路为: 从用户的行为序列中学习一个用户上下文变量, 以此来推断每个用户的偏好, 从而使推荐系统更好地适应仅有少量交互的新用户; 从信息论的角度出发, 引入潜在策略空间中的互信息正则化来建模用户模型与推荐系统之间的互动关系, 从而实现对冷启动用户的快速适应; 为提高自适应效率, 利用逆强化学习方法从少量交互中恢复出用户的策略与奖励, 从而辅助推荐系统. 在模拟与真实的推荐数据集中的结果证明了 M^{3}Rec 的有效性. 所提算法不仅适用于推荐系统, 还能推广至其他领域, 如机器人控制.

同样为缓解数据稀疏问题并提高推荐性能, Nie 等[60] 从奖励函数和状态表征两个角度对离线强化学习框架进行改进, 并提出一种基于知识增强的因果强化学习模型 (Knowledge-enhanced causal reinforcement learning, KCRL). 该模型分为用户模型学习和强化学习策略学习两个阶段. 在用户模型学习阶段, 为优化推荐策略学习中交付的用户满意度 (即奖励), 构建一个群体偏好注入的因果用户模型 (Group preference-injected causal user model, GCUM), 采用因果推理来模拟真实用户偏好信息. 在强化学习策略学习阶段, 训练一个强化学习策略模块来学习 GCUM 提供的因果用户满意度 (即奖励) 的推荐策略. 为进一步减轻数据稀疏性, 设计知识增强状态编码器来丰富用户状态表征. 在真实的短视频与电影推荐数据集上进行测试, 实验结果表明, 与现有的推荐方法相比, KCRL 可以更好地缓解数据稀疏性问题, 并获得更好的推荐性能.

基于文本的交互式推荐系统是指运用自然语言处理技术来理解用户的意图和需求, 并在推荐过程中允许用户提供反馈, 从而持续优化推荐结果, 为用户提供更加准确、个性化的服务. 然而, 在离线环境中, 系统无法直接与用户进行交互, 而是从多个未知策略收集的经验数据中学习, 因此很容易受到分布偏移的影响. 为解决该问题, Zhang 等[61] 提出一个行为不可知的偏离策略校正框架 OIR (Offline interactive recommendation), 实现在离线环境下进行交互式推荐. 首先构建多模态编码器, 将用户的自然语言反馈数据 (文本) 与候选产品图像分别进行编码, 得到各自的表征, 再将两者共同作为多层感知机的输入. 多模态编码器的输出结果作为离线强化学习的输入, 利用保守的 Q 函数来执行异策略评估. 所提算法能从固定的数据集中学习有效的策略, 而无需进一步的交互. 在 UT-Zappos 50K 环境上测试, 结果表明, 相比于基准算法, 所提框架能够更准确地估计奖励校正量. 此外, 离线训练方案在离线交互推荐系统中表现出优于基线的性能.

传统的在线交互模式需要不断与用户进行交互来收集经验, 然而这在很多现实场景中并不可行. 例如, 安全关键系统中, 在性能和安全性得到验证之前, 学到的策略不应该被应用于实际系统. 离线交互推荐则是一种基于历史数据进行计算和分析, 生成用户推荐结果的离线推荐方式. 这种方式不需要实时计算, 因此可以更快地响应用户的请求, 并保护用户的隐私. 另外, 由于可以利用用户的历史行为数据来进行推荐, 所以能更准确地预测用户的兴趣和偏好. 因此, 将离线强化学习应用于交互推荐领域是一个非常有前景的解决方案, 可以扩展到各种领域, 如智能终端应用程序、电子商务平台、社交网络、医疗诊断等, 为用户提供更加个性化和准确的推荐服务.

4.3 智能驾驶

随着科技的进步, 驾驶模拟器已逐渐成为研究和开发智能驾驶技术的重要工具之一. 这些模拟器能够生成海量的驾驶场景数据, 这对于训练基于学习的车辆规划算法来说至关重要. 然而, 在处理这些数据时, 传统的模仿学习方法受到多种限制, 其中最核心的问题是专家数据质量以及协变量移位现象. 相比之下, 离线强化学习可以利用大规模实际离线驾驶数据进行训练, 克服了专家数据质量的限制. 其次, 通过使用大规模离线数据集中的多样化示例来训练模型, 离线强化学习能够更好地适应新的环境, 解决协变量移位问题. 此外, 通过在离线数

据集中学习和优化策略, 离线强化学习能够减少在实际驾驶中的风险探索, 从而提供更安全的驾驶决策.

在实际应用中, 与环境交互来更新策略可能导致智能驾驶汽车发生严重的交通事故. 为此, 张健等[58] 通过 SUMO 仿真软件来模拟交叉口场景下车辆的通行控制过程, 在此基础上, 将基于表征学习的离线强化学习算法 BCQ 成功应用于解决城市工况智能网联车辆信号交叉口场景下的生态驾驶问题. 仿真结果表明, BCQ 算法能改善智能网联车辆的燃油经济性, 实现节能控制. 这些成功应用也进一步证实离线强化学习能为智能驾驶领域提供切实可行的指导方案.

将离线强化学习算法应用于智能驾驶领域时, 可能会遇到长视域问题. 在长视域任务中, 奖励信号往往稀疏且具有延迟性, 智能体需要从中学习并规划合理的动作序列, 以避免在较长时间内出现复合误差. 为应对这一挑战, Li 等[69] 提出一种基于分层技能的车辆规划离线强化学习 (Hierarchical skill-based offline reinforcement learning for vehicle planning, HsO-VP) 方法. 该算法首先利用变分自编码器从离线示范数据中学习技能. 该编码器采用双分支网络架构, 其中离散分支用于区分不同的驾驶技能选项, 而连续分支则捕获特定技能选项内的变化, 从而缓解模型后验坍塌问题. 在此基础上, 训练一个高阶策略, 该策略输出的是技能而不是每个时间步的动作, 因此可以实现对未来的长期推理. 在 CARLA 仿真平台上的综合实验证明 HsO-VP 算法具有良好的泛化性能和鲁棒性, 能够应对复杂的驾驶场景和长视域的离线学习任务.

总体来说, 基于表征学习的离线强化学习是一种有效的策略学习方法, 它利用大规模离线数据集克服传统模仿学习的限制, 并避免了真实驾驶环境中的风险探索.

除上述应用外, 在医疗领域, PerSim[63] 算法通过利用先前其他患者的临床病史数据来为新患者学习个性化的模拟器, 从而提供准确的个性化治疗方案; 在能源管理领域, He 等[59] 提出基于 BCQ 的网联油电混动汽车能源管理策略, 旨在提高网联汽车在复杂行驶工况下的能源利用效率和环境适应性; 在量化交易领域, ORDC (Offline RL with dynamic context) 算法[70] 采用紧凑的环境表征学习方法, 有效地缓解了离线强化学习在交易执行任务上的过拟合问题.

然而, 将基于表征学习的离线强化学习用于现实场景仍面临一些挑战: 收集到的数据可能不够全面或无法覆盖所有可能的场景, 从而导致模型训练不足或过拟合; 如何在保证安全性的前提下, 得到一个可行的最优策略; 如何提高模型的泛化能力和适应性; 现有的传感器数据往往是部分可观测的或者不确定的情况; 奖励信号的设计难度大; 如何选择和设计适当的表征学习方法; 模型训练时间长, 需要大量的计算资源等. 尽管如此, 我们相信随着仿真平台和大规模离线数据集的发展, 基于表征学习的离线强化学习在现实场景的应用前景将会更加广阔.

5 总结与展望

本文从方法、数据、评估与应用四大层面对近几年来基于表征学习的离线强化学习方法研究进展进行全面概述. 首先对离线强化学习问题进行形式化描述, 将离线设置下难以学到最优策略的原因归结为外推误差, 而导致外推误差的两个重要因素为分布偏移与数据覆盖不足. 在此基础上, 针对这两个问题并根据表征对象的不同, 将现有的基于表征学习的离线强化学习方法分为动作表征、状态表征、状态−动作对表征、轨迹表征以及任务或环境表征五类, 并详细分析了每种类别下的典型算法. 在实验设置层面, 为公平比较离线强化学习算法性能, 介绍了 3 种基准数据集与评估协议, 同时对现有的离线策略评估与超参数选择的研究进展进行回顾与总结. 最后, 给出了离线强化学习在工业、推荐系统、智能驾驶等领域的应用.

尽管基于表征学习的离线强化学习领域已经取得一定的进展, 但还有许多挑战和问题需要进一步研究和解决.

1) 数据层面. 现有方法通过数据增强、生成式模型、Pareto 策略池等方式来增强离线数据的多样性. 能否有更好的自监督数据处理方式来解决数据覆盖不足的问题, 是未来值得探讨的问题.

2) 理论层面. 当前的理论研究大都局限于表格马尔科夫决策过程与线性马尔科夫决策过程. 针对非线性马尔科夫决策过程, 证明的基本思想大都采用近似化手段, 将非线性问题通过函数映射转化为线性问题求解, 能否有更好的手段在一般马尔科夫决策过程上证明离线强化学习的采样效率, 是离线强化学习理论研究的难点之一.

3) 方法层面. 当前, 离线强化学习算法在处理复杂、连续和高维度状态与动作时仍面临诸多挑战. 如何更有效地从数据中提取表征以准确表示状态和动作是一个关键问题. 由于现实世界的多样性和复杂性, 仅凭有限数据构建通用表征十分困难. 尽管

扩散模型能够学习到良好的表征, 但其训练过程效率低下, 无法直接应用于实际问题. 因此, 有必要探索更高效的方法和技术来增强表征学习能力, 并提高其泛化性能. 同时, 基于表征学习的离线强化学习算法的安全性和可解释性也是一个重要的挑战. 由于这些算法是基于数据学习的, 可能会产生不可预测的行为和决策. 因此, 有必要研究更安全、更可解释的算法, 以确保智能决策的可靠性和透明性.

4) 离线策略评估与超参数选择. 由于无法与环境交互, 模型的超参数很难调整, 导致实际选择时倾向于过于保守的策略. 因此设计合理有效的离线策略评估 (选择) 协议也是该方向亟需解决的问题.

相信随着研究的不断深入, 离线强化学习这种数据驱动的方式能够真正在实际应用场景中落地, 解决更多现实领域的决策控制难题.

References

1 Sutton R S, Barto A G. *Reinforcement Learning: An Introduction* (Second edition). Cambridge: The MIT Press, 2018.

2 Sun Yue-Wen, Liu Wen-Zhang, Sun Chang-Yin. Causality in reinforcement learning control: The state of the art and prospects. *Acta Automatica Sinica*, 2023, **49**(3): 661−677
(孙悦雯, 柳文章, 孙长银. 基于因果建模的强化学习控制: 现状及展望. 自动化学报, 2023, **49**(3): 661−677)

3 Silver D, Huang A, Maddison C J, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016, **529**(7587): 484−489

4 Schrittwieser J, Antonoglou I, Hubert T, Simonyan K, Sifre L, Schmitt S, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 2020, **588**(7839): 604−609

5 Senior A W, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 2020, **577**(7792): 706−710

6 Li Y J, Choi D, Chung J, Kushman N, Schrittwieser J, Leblond R, et al. Competition-level code generation with AlphaCode. *Science*, 2022, **378**(6624): 1092−1097

7 Degrave J, Felici F, Buchli J, Neunert M, Tracey B, Carpanese F, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 2022, **602**(7897): 414−419

8 Fawzi A, Balog M, Huang A, Hubert T, Romera-Paredes B, Barekatain M, et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 2022, **610**(7930): 47−53

9 Fang X, Zhang Q C, Gao Y F, Zhao D B. Offline reinforcement learning for autonomous driving with real world driving data. In: Proceedings of the 25th IEEE International Conference on Intelligent Transportation Systems (ITSC). Macao, China: IEEE, 2022. 3417−3422

10 Liu Jian, Gu Yang, Cheng Yu-Hu, Wang Xue-Song. Prediction of breast cancer pathogenic genes based on multi-agent reinforcement learning. *Acta Automatica Sinica*, 2022, **48**(5): 1246−1258
(刘健, 顾扬, 程玉虎, 王雪松. 基于多智能体强化学习的乳腺癌致病基因预测. 自动化学报, 2022, **48**(5): 1246−1258)

11 Levine S, Kumar A, Tucker G, Fu J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv preprint arXiv: 2005.01643, 2020.

12 Prudencio R F, Maximo M R O A, Colombini E L. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, DOI: 10.1109/TNNLS.2023.3250269

13 Cheng Yu-Hu, Huang Long-Yang, Hou Di-Yuan, Zhang Jia-Zhi, Chen Jun-Long, Wang Xue-Song. Generalized offline actor-critic with behavior regularization. *Chinese Journal of Computers*, 2023, **46**(4): 843−855
(程玉虎, 黄龙阳, 侯棣元, 张佳志, 陈俊龙, 王雪松. 广义行为正则化离线 Actor-Critic. 计算机学报, 2023, **46**(4): 843−855)

14 Gu Yang, Cheng Yu-Hu, Wang Xue-Song. Offline reinforcement learning based on prioritized sampling model. *Acta Automatica Sinica*, 2024, **50**(1): 143−153
(顾扬, 程玉虎, 王雪松. 基于优先采样模型的离线强化学习. 自动化学报, 2024, **50**(1): 143−153)

15 Fujimoto S, Meger D, Precup D. Off-policy deep reinforcement learning without exploration. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR, 2019. 2052−2062

16 He Q, Hou X W, Liu Y. POPO: Pessimistic offline policy optimization. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022. 4008−4012

17 Wu J L, Wu H X, Qiu Z H, Wang J M, Long M S. Supported policy optimization for offline reinforcement learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 2268

18 Lyu J F, Ma X T, Li X, Lu Z Q. Mildly conservative *Q*-learning for offline reinforcement learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 125

19 Rezaeifar S, Dadashi R, Vieillard N, Hussenot L, Bachem O, Pietquin O, et al. Offline reinforcement learning as anti-exploration. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtual Event: AAAI Press, 2022. 8106−8114

20 Zhou W X, Bajracharya S, Held D. PLAS: Latent action space for offline reinforcement learning. In: Proceedings of the 4th Conference on Robot Learning. Cambridge, USA: PMLR, 2020. 1719−1735

21 Chen X, Ghadirzadeh A, Yu T H, Wang J H, Gao A, Li W Z, et al. LAPO: Latent-variable advantage-weighted policy optimization for offline reinforcement learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 2674

22 Akimov D, Kurenkov V, Nikulin A, Tarasov D, Kolesnikov S. Let offline RL flow: Training conservative agents in the latent space of normalizing flows. In: Proceedings of Offline Reinforcement Learning Workshop at Neural Information Processing Systems. New Orleans, USA: OpenReview.net, 2022.

23 Yang Y Q, Hu H, Li W Z, Li S Y, Yang J, Zhao Q C, et al. Flow to control: Offline reinforcement learning with lossless primitive discovery. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press, 2023. 10843−10851

24 Wang Z D, Hunt J J, Zhou M Y. Diffusion policies as an expressive policy class for offline reinforcement learning. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net, 2023.

25 Chen H Y, Lu C, Ying C Y, Su H, Zhu J. Offline reinforcement learning via high-fidelity generative behavior modeling. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net, 2023.

26 Zhang H C, Shao J Z, Jiang Y H, He S C, Zhang G W, Ji X Y. State deviation correction for offline reinforcement learning. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtual Event: AAAI Press, 2022. 9022−9030

27 Weissenbacher M, Sinha S, Garg A, Kawahara Y. Koopman Q-learning: Offline reinforcement learning via symmetries of dynamics. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 23645−23667

28 Rafailov R, Yu T H, Rajeswaran A, Finn C. Offline reinforcement learning from images with latent space models. In: Pro-

ceedings of the 3rd Annual Conference on Learning for Dynamics and Control. Zurich, Switzerland: PMLR, 2021. 1154−1168

29 Cho D, Shim D, Kim H J. S2P: State-conditioned image synthesis for data augmentation in offline reinforcement learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 838

30 Gieselmann R, Pokorny F T. An expansive latent planner for long-horizon visual offline reinforcement learning. In: Proceedings of the RSS 2023 Workshop on Learning for Task and Motion Planning. Daegu, South Korea: OpenReview.net, 2023.

31 Zang H Y, Li X, Yu J, Liu C, Islam R, Combes R T D, et al. Behavior prior representation learning for offline reinforcement learning. In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net, 2023.

32 Mazoure B, Kostrikov I, Nachum O, Tompson J. Improving zero-shot generalization in offline reinforcement learning using generalized similarity functions. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 1819

33 Kim B, Oh M H. Model-based offline reinforcement learning with count-based conservatism. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR, 2023. 16728−16746

34 Tennenholtz G, Mannor S. Uncertainty estimation using riemannian model dynamics for offline reinforcement learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 1381

35 Ada S E, Oztop E, Ugur E. Diffusion policies for out-of-distribution generalization in offline reinforcement learning. *IEEE Robotics and Automation Letters*, 2024, **9**(4): 3116−3123

36 Kumar A, Agarwal R, Ma T Y, Courville A C, Tucker G, Levine S. DR3: Value-based deep reinforcement learning requires explicit regularization. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: OpenReview.net, 2022.

37 Lee B J, Lee J, Kim K E. Representation balancing offline model-based reinforcement learning. In: Proceedings of the 9th International Conference on Learning Representations. Virtual Event: OpenReview.net, 2021.

38 Chang J D, Wang K W, Kallus N, Sun W. Learning bellman complete representations for offline policy evaluation. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 2938−2971

39 Chen L L, Lu K, Rajeswaran A, Lee K, Grover A, Laskin M, et al. Decision transformer: Reinforcement learning via sequence modeling. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates, Inc., 2021. 15084−15097

40 Janner M, Li Q Y, Levine S. Offline reinforcement learning as one big sequence modeling problem. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates, Inc., 2021. 1273−1286

41 Furuta H, Matsuo Y, Gu S S. Generalized decision transformer for offline hindsight information matching. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: OpenReview.net, 2022.

42 Liu Z X, Guo Z J, Yao Y H, Cen Z P, Yu W H, Zhang T N, et al. Constrained decision transformer for offline safe reinforcement learning. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org, 2023. Article No. 893

43 Wang Y Q, Xu M D, Shi L X, Chi Y J. A trajectory is worth three sentences: Multimodal transformer for offline reinforcement learning. In: Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence. Pittsburgh, USA: JMLR.org, 2023. Article No. 208

44 Zeng Z L, Zhang C, Wang S J, Sun C. Goal-conditioned predictive coding for offline reinforcement learning. arXiv preprint arXiv: 2307.03406, 2023.

45 Janner M, Du Y L, Tenenbaum J B, Levine S. Planning with diffusion for flexible behavior synthesis. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 9902−9915

46 Ajay A, Du Y L, Gupta A, Tenenbaum J B, Jaakkola T S, Agrawal P. Is conditional generative modeling all you need for decision making? In: Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net, 2023.

47 Liang Z X, Mu Y, Ding M Y, Ni F, Tomizuka M, Luo P. AdaptDiffuser: Diffusion models as adaptive self-evolving planners. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org, 2023. Article No. 854

48 Yuan H Q, Lu Z Q. Robust task representations for offline meta-reinforcement learning via contrastive learning. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 25747−25759

49 Zhao C Y, Zhou Z H, Liu B. On context distribution shift in task representation learning for online meta RL. In: Proceedings of the 19th Advanced Intelligent Computing Technology and Applications. Zhengzhou, China: Springer, 2023. 614−628

50 Chen X H, Yu Y, Li Q Y, Luo F M, Qin Z W, Shang W J, et al. Offline model-based adaptable policy learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates, Inc., 2021. 8432−8443

51 Sang T, Tang H Y, Ma Y, Hao J Y, Zheng Y, Meng Z P, et al. PAnDR: Fast adaptation to new environments from offline experiences via decoupling policy and environment representations. In: Proceedings of the 31st International Joint Conference on Artificial Intelligence. Vienna, Austria: IJCAI, 2022. 3416−3422

52 Lou X Z, Yin Q Y, Zhang J G, Yu C, He Z F, Cheng N J, et al. Offline reinforcement learning with representations for actions. *Information Sciences*, 2022, **610**: 746−758

53 Kingma D P, Welling M. Auto-encoding variational Bayes. In: Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada: ICLR, 2014.

54 Mark M S, Ghadirzadeh A, Chen X, Finn C. Fine-tuning offline policies with optimistic action selection. In: Proceedings of NeurIPS Workshop on Deep Reinforcement Learning. Virtual Event: OpenReview.net, 2022.

55 Zhang Bo-Wei, Zheng Jian-Fei, Hu Chang-Hua, Pei Hong, Dong Qing. Missing data generation method based on flow model and its application in remaining life prediction. *Acta Automatica Sinica*, 2023, **49**(1): 185−196
(张博玮, 郑建飞, 胡昌华, 裴洪, 董青. 基于流模型的缺失数据生成方法在剩余寿命预测中的应用. 自动化学报, 2023, **49**(1): 185−196)

56 Yang L, Zhang Z L, Song Y, Hong S D, Xu R S, Zhao Y, et al. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 2023, **56**(4): Article No. 105

57 Wang Y N, Ge Y, Li L, Chen R, Xu T. Offline meta-level model-based reinforcement learning approach for cold-start recommendation. arXiv preprint arXiv: 2012.02476, 2020.

58 Zhang Jian, Jiang Xia, Shi Xiao-Yu, Cheng Jian, Zheng Yue-Biao. Offline reinforcement learning for eco-driving control at signalized intersections. *Journal of Southeast University (Natural Science Edition)*, 2022, **52**(4): 762−769
(张健, 姜夏, 史晓宇, 程健, 郑岳标. 基于离线强化学习的交叉口生态驾驶控制. 东南大学学报 (自然科学版), 2022, **52**(4): 762−769)

59 He H W, Niu Z G, Wang Y, Huang R C, Shou Y W. Energy management optimization for connected hybrid electric vehicle using offline reinforcement learning. *Journal of Energy Storage*, 2023, **72**: Article No. 108517

60 Nie W Z, Wen X, Liu J, Chen J W, Wu J C, Jin G Q, et al. Knowledge-enhanced causal reinforcement learning model for interactive recommendation. *IEEE Transactions on Multimedia*, 2024, **26**: 1129−1142

61 Zhang R Y, Yu T, Shen Y L, Jin H Z. Text-based interactive recommendation via offline reinforcement learning. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtual Event: AAAI Press, 2022. 11694−11702

62 Rigter M, Lacerda B, Hawes N. RAMBO-RL: Robust adversarial model-based offline reinforcement learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. 16082−16097

63 Agarwal A, Alomar A, Alumootil V, Shah D, Shen D, Xu Z, et al. PerSim: Data-efficient offline reinforcement learning with heterogeneous agents via personalized simulators. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates, Inc., 2021. 18564−18576

64 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017. 6000−6010

65 Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. An image is worth 16×16 words: Transformers for image recognition at scale. In: Proceedings of the 9th International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2021.

66 Wang Xue-Song, Wang Rong-Rong, Cheng Yu-Hu. Safe reinforcement learning: A survey. *Acta Automatica Sinica*, 2023, **49**(9): 1813−1835
(王雪松, 王荣荣, 程玉虎. 安全强化学习综述. 自动化学报, 2023, **49**(9): 1813−1835)

67 Lai Y, Liu J X, Tang Z T, Wang B, Hao J Y, Luo P. ChiPFormer: Transferable chip placement via offline decision transformer. In: Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: PMLR, 2023. 18346−18364

68 Zhao T Z, Luo J L, Sushkov O, Pevceviciute R, Heess N, Scholz J, et al. Offline meta-reinforcement learning for industrial insertion. In: Proceedings of International Conference on Robotics and Automation. Philadelphia, USA: IEEE, 2022. 6386−6393

69 Li Z N, Nie F, Sun Q, Da F, Zhao H. Boosting offline reinforcement learning for autonomous driving with hierarchical latent skills. arXiv preprint arXiv: 2309.13614, 2023.

70 Zhang C H, Duan Y T, Chen X Y, Chen J Y, Li J, Zhao L. Towards generalizable reinforcement learning for trade execution. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macao, China: IJCAI, 2023. Article No. 553

71 Gulcehre C, Wang Z Y, Novikov A, Le Paine T, Colmenarejo S G, Zołna K, et al. RL unplugged: A suite of benchmarks for offline reinforcement learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 608

72 Fu J, Kumar A, Nachum O, Tucker G, Levine S. D4RL: Datasets for deep data-driven reinforcement learning. arXiv preprint arXiv: 2004.07219, 2020.

73 Qin R J, Zhang X Y, Gao S Y, Chen X H, Li Z W, Zhang W N, et al. NeoRL: A near real-world benchmark for offline reinforcement learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 1795

74 Song H F, Abdolmaleki A, Springenberg J T, Clark A, Soyer H, Rae J W, et al. V-MPO: On-policy maximum a posteriori policy optimization for discrete and continuous control. In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: Open Review.net, 2020.

75 Merel J, Hasenclever L, Galashov A, Ahuja A, Pham V, Wayne G, et al. Neural probabilistic motor primitives for humanoid control. In: Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net, 2019.

76 Merel J, Aldarondo D, Marshall J, Tassa Y, Wayne G, Olveczky B. Deep neuroethology of a virtual rodent. In: Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: OpenReview.net, 2020.

77 Machado M C, Bellemare M G, Talvitie E, Veness J, Hausknecht M, Bowling M. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 2018, **61**: 523−562

78 Dulac-Arnold G, Levine N, Mankowitz D J, Li J, Paduraru C, Gowal S, et al. An empirical investigation of the challenges of real-world reinforcement learning. arXiv preprint arXiv: 2003.11881, 2020.

79 Abdolmaleki A, Springenberg J T, Tassa Y, Munos R, Heess N, Riedmiller M A. Maximum a posteriori policy optimisation. In: Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net, 2018.

80 Pomerleau D A. ALVINN: An autonomous land vehicle in a neural network. In: Proceedings of the 1st International Conference on Neural Information Processing Systems. Denver, USA: MIT Press, 1988. 305−313

81 Mnih V, Kavukcuoglu K, Silver D, Rusu A A, Veness J, Bellemare M G, et al. Human-level control through deep reinforcement learning. *Nature*, 2015, **518**(7540): 529−533

82 Barth-Maron G, Hoffman M W, Budden D, Dabney W, Horgan D, Dhruva T B, et al. Distributed distributional deterministic policy gradients. In: Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net, 2018.

83 Dabney W, Ostrovski G, Silver D, Munos R. Implicit quantile networks for distributional reinforcement learning. In: Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018. 1104−1113

84 Wu Y F, Tucker G, Nachum O. Behavior regularized offline reinforcement learning. arXiv preprint arXiv: 1911.11361, 2019.

85 Siegel N, Springenberg J T, Berkenkamp F, Abdolmaleki A, Neunert M, Lampe T, et al. Keep doing what worked: Behavior modelling priors for offline reinforcement learning. In: Proceedings of International Conference on Learning Representations. Addis Ababa, Ethiopia: OpenReview.net, 2020.

86 Agarwal A, Schuurmans D, Norouzi M. An optimistic perspective on offline reinforcement learning. In: Proceedings of the 37th International Conference on Machine Learning. Virtual Event: PMLR, 2020. 104−114

87 Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018. 1856−1865

88 Kumar A, Fu J, Soh M, Tucker G, Levine S. Stabilizing off-policy Q-learning via bootstrapping error reduction. In: Proceedings of the International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc., 2019. 11761−11771

89 Peng X B, Kumar A, Zhang G, Levine S. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. arXiv preprint arXiv: 1910.00177, 2019.

90 Kumar A, Zhou A, Tucker G, Levine S. Conservative Q-learning for offline reinforcement learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 100

91 Nachum O, Dai B, Kostrikov I, Chow Y, Li L H, Schuurmans D. AlgaeDICE: Policy gradient from arbitrary experience. arXiv preprint arXiv: 1912.02074, 2019.

92 Wang Z Y, Novikov A, Żołna K, Springenberg J T, Reed S, Shahriari B, et al. Critic regularized regression. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 651

93 Matsushima T, Furuta H, Matsuo Y, Nachum O, Gu S X. Deployment-efficient reinforcement learning via model-based offline optimization. In: Proceedings of the 9th International Conference on Learning Representations. Virtual Event: OpenReview.net, 2021.

94 Yu T H, Thomas G, Yu L T, Ermon S, Zou J, Levine S, et al. MOPO: Model-based offline policy optimization. In: Proceedings of the 34th International Conference on Advances in Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 1185

95 Le H M, Voloshin C, Yue Y S. Batch policy learning under constraints. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR, 2019. 3703−3712

96 Koller D, Friedman N. *Probabilistic Graphical Models: Principles and Techniques*. Cambridge: MIT Press, 2009.

97 Wang Shuo-Ru, Niu Wen-Jia, Tong En-Dong, Chen Tong, Li He, Tian Yun-Zhe, et al. Research on off-policy evaluation in reinforcement learning: A survey. *Chinese Journal of Computers*, 2022, **45**(9): 1926−1945
(王硕汝, 牛温佳, 童恩栋, 陈彤, 李赫, 田蕴哲, 等. 强化学习离线策略评估研究综述. 计算机学报, 2022, **45**(9): 1926−1945)

98 Fu J, Norouzi M, Nachum O, Tucker G, Wang Z Y, Novikov A, et al. Benchmarks for deep off-policy evaluation. In: Proceedings of the 9th International Conference on Learning Representations. Virtual Event: OpenReview.net, 2021.

99 Schweighofer K, Dinu M, Radler A, Hofmarcher M, Patil V P, Bitto-nemling A, et al. A dataset perspective on offline reinforcement learning. In: Proceedings of the 1st Conference on Lifelong Learning Agents. McGill University, Canada: PMLR, 2022. 470−517

100 Konyushkova K, Chen Y T, Paine T, Gülçehre C, Paduraru C, Mankowitz D J, et al. Active offline policy selection. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates, Inc., 2021. 24631−24644

101 Kurenkov V, Kolesnikov S. Showing your offline reinforcement learning work: Online evaluation budget matters. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 11729−11752

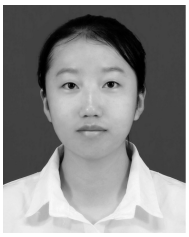
102 Lu C, Ball P J, Parker-Holder J, Osborne M A, Roberts S J. Revisiting design choices in offline model based reinforcement learning. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: OpenReview.net, 2022.

103 Hu H, Yang Y Q, Zhao Q C, Zhang C J. On the role of discount factor in offline reinforcement learning. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 9072−9098

104 Nair A, Zhu B, Narayanan G, Solowjow E, Levine S. Learning on the job: Self-rewarding offline-to-online finetuning for industrial insertion of novel connectors from vision. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). London, United Kingdom: The IEEE, 2023. 7154−7161

105 Kostrikov I, Nair A, Levine S. Offline reinforcement learning with implicit Q-learning. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: OpenReview.net, 2022.

王雪松 中国矿业大学信息与控制工程学院教授. 2002 年获得中国矿业大学博士学位. 主要研究方向为机器学习与模式识别.
E-mail: wangxuesongcumt@163.com
(**WANG Xue-Song** Professor at the School of Information and Control Engineering, China University of Mining and Technology. She received her Ph.D. degree from China University of Mining and Technology in 2002. Her research interest covers machine learning and pattern recognition.)



王荣荣 中国矿业大学信息与控制工程学院博士研究生. 2021 年获得济南大学硕士学位. 主要研究方向为深度强化学习.
E-mail: wangrongrong1996@126.com
(**WANG Rong-Rong** Ph.D. candidate at the School of Information and Control Engineering, China University of Mining and Technology. She received her master degree from University of Jinan in 2021. Her main research interest is deep reinforcement learning.)

程玉虎 中国矿业大学信息与控制工程学院教授. 2005 年获得中国科学院自动化研究所博士学位. 主要研究方向为机器学习与智能系统. 本文通信作者. E-mail: chengyuhu@163.com
(**CHENG Yu-Hu** Professor at the School of Information and Control Engineering, China University of Mining and Technology. He received his Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences in 2005. His research interest covers machine learning and intelligent system. Corresponding author of this paper.)