离线强化学习研究综述

乌 兰 刘 全 黄志刚 张立华

(苏州大学计算机科学与技术学院 江苏 苏州 215006)

摘 要 离线强化学习也称为批量强化学习,是深度强化学习领域的一项重要研究内容。它利用行为策略生成静态数据集,无需在线和环境交互,成功地将大规模数据集转变成强大的决策引擎。近年来,离线强化学习方法得到了广泛关注和深入研究,并在实际应用中取得了瞩目的成绩。目前,该方法已经用于推荐系统、导航驾驶、自然语言处理、机器人控制以及医疗与能源等应用领域,并被看作是现实世界应用强化学习最具潜力的技术途径之一。该文首先介绍了离线强化学习的背景与理论基础。随后从求解思路出发,将离线强化学习方法分为无模型、基于模型和基于 Transformer 模型 3 大类,并对各类方法的研究现状与发展趋势进行分析。同时,对比了目前 3 个最流行的实验环境 D4RL、RL Unplugged 和 NeoRL。进而介绍了离线强化学习技术在现实世界诸多领域的应用。最后,对离线强化学习进行总结与展望,以此推动更多该领域的研究工作。

关键词 人工智能;强化学习;深度强化学习;离线强化学习;批量强化学习

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2025.00156

A Review of Research on Offline Reinforcement Learning

WU Lan LIU Quan HUANG Zhi-Gang ZHANG Li-Hua

(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006)

Abstract Batch Reinforcement Learning is an important branch in the field of reinforcement learning. As the need to rely on historical data for reinforcement learning became more and more pressing, offline reinforcement learning was not systematically proposed until 2020. Therefore, offline reinforcement learning, also known as batch reinforcement learning, is an important research topic in the field of deep reinforcement learning. By utilizing behavior policies to generate static datasets and without online interaction with the environment, this approach successfully converts large datasets into powerful decision engines. The rise of offline reinforcement learning has not only accelerated the development of decision engines but also provided researchers with a stable and efficient training framework. In recent years, offline reinforcement learning methods have received extensive attention and have undergone in-depth research, achieving remarkable results in practical applications. Currently, these methods have been used in recommendation systems, navigation, driving, natural language processing, and robot control, as well as in the fields of healthcare and energy, and are considered one of the most promising technology approaches for applying reinforcement learning in the real world. In this paper, we first introduce the background and theoretical basis of offline reinforcement learning. Secondly, starting from the solution idea,

收稿日期:2023-09-06;在线发布日期:2024-06-21。本课题得到国家自然科学基金(62376179,62176175)、新疆维吾尔自治区自然科学基金(2022D01A238)、江苏高校优势学科建设工程资助项目(PAPD)资助。**乌 兰**,博士研究生,中国计算机学会(CCF)会员,主要研究方向为深度强化学习、离线强化学习。E-mail:20217927001@stu.suda.edu.cn。**刘 全**(通信作者),博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为深度强化学习、自动推理。E-mail:quanliu@suda.edu.cn。**黄志刚**,博士研究生,中国计算机学会(CCF)会员,主要研究方向为深度强化学习、分层强化学习。**张立华**,博士研究生,中国计算机学会(CCF)会员,主要研究方向为深度强化学习、逆向强化学习。

the offline reinforcement learning methods are classified into three major categories: model-free, model-based, and transformer-based. In the meantime, we analyze the research status and development trends of each method. Specifically, these methods do not share the same focus and aim to address distinct challenges, achieving incremental improvements in handling distribution shifts. Model-free offline reinforcement learning methods focus on policy evaluation and improvement by directly utilizing trajectory information from static data. In contrast, model-based offline reinforcement learning methods aim to learn dynamic environment models from static datasets to optimize policies. Recently, transformer-based offline reinforcement learning methods have attracted prominence due to their superior sequence modeling abilities, showing exceptional performance in managing complex environments and long-term sequential data. Thirdly, we compare the three most popular experimental environments D4RL, RL Unplugged, and NeoRL. They offer rich datasets and standardized evaluation metrics to compare the effectiveness and stability of various offline reinforcement learning algorithms. D4RL and RL Unplugged are biased towards simulation platforms, while NeoRL is biased towards practical applications. Specifically, D4RL includes navigation, manipulation, and locomotion tasks. RL Unplugged includes manipulation, locomotion, and game tasks. NeoRL includes industrial benchmarking, a stock exchange simulator, and city management tasks. Then, we introduce the applications of offline reinforcement learning in multiple real-world fields. These applications demonstrate the potential and value of offline reinforcement learning in solving real-world problems. Finally, we provide prospects and summaries for offline reinforcement learning, to promote more research in this field. In the future, with a deeper understanding of the theory of offline reinforcement learning and further technological advancements, it is anticipated that this field will continue to attract increasing research attention. Offline reinforcement learning combines the advantages of deep learning and reinforcement learning and is expected to provide smarter and more efficient solutions to various complex tasks.

Keywords　artificial intelligence; reinforcement learning; deep reinforcement learning; offline reinforcement learning; batch reinforcement learning

1 引　言

强化学习[1](Reinforcement Learning,RL)作为机器学习领域中的一个重要研究方向,它把生物学中的试错方式与数学中的最优控制问题结合起来,是一种交互式学习方法。深度强化学习[2](Deep Reinforcement Learning,DRL)打破了传统智能算法的研究,创新性地将具有感知能力的深度学习[3](Deep Learning,DL)与具有决策能力的强化学习相结合,形成一种端到端的完整智能系统。谷歌的人工智能研究团队 DeepMind 将 RL 与蒙特卡洛树搜索方法相结合[4],开发出了首次击败人类顶尖职业选手的围棋系统 AlphaGo[5],这成为人工智能领域的重大突破。DRL 在虚拟游戏等环境中已经展现出超越人类专家的性能。同样地,现实世界场景也对 DRL 提出了迫切需求,它的应用落地可以产生重大价值。然而,DRL 需要智能体与环境在线交互,并且进行大量数据采样。在许多情况下,这种在线互动是不切实际的。因为在现实世界中的代价昂贵且试错风险极高,由此产生的成本与安全问题不容忽视。

批量强化学习[6-8](Batch Reinforcement Learning,BRL)是 RL 领域的一个重要分支。随着依赖历史数据进行 RL 的需求越来越迫切,直到 2020 年,Levine 等人[9]从 BRL 概念引申而来,系统地提出了离线强化学习(Offline Reinforcement Learning,Offline RL)。由此正式阐明了该学习所面对的挑战,并初步提出了建设性的解决方案。离线 RL 作为在现实世界应用中最有潜力的技术之一,它由未知的行为策略产生离线数据,并利用这些数据进行当前策略的学习与更新[10]。事实上,数据驱动(Data-Driven)在解决机器学习的实际问题中具有不可或缺的作用。对于 DRL,这一问题启发了类似的思考:是否可以针对 RL 目标应用同样的数据驱动学习,从而建立

数据驱动的 RL 框架。与传统在线(Online)RL 不同,离线 RL 只利用先前收集的离线数据来构造静态数据集,而不与环境进行额外的在线交互。离线 RL 的出现给 RL 应用落地提供了挑战与机遇。一方面,依赖静态数据集的离线 RL 引发了诸多问题。在线异策略(Online Off-Policy)RL 可转化成离线 RL,但是智能体无法与环境进行交互,从而导致这些方法的性能下降。面对高维空间和函数逼近的情况下,算法在处理环境中的动态变化时面临困境。其根本原因在于函数逼近使得算法更容易受到数据分布变化的影响,从而增加了算法对数据分布变动的敏感性。由此产生的分布偏移(Distribution Shift)问题,正是离线 RL 的核心挑战之一。另一方面,离线 RL 带来了新的机遇。它通过利用数据集学习策略,避免了交互产生的成本损失,并提高了 RL 技术在现实世界中的安全性。

离线 RL 作为新提出的 DRL 方法,已经在短时间受到了人工智能领域的高度关注。目前,最简单的方法是通过限制当前策略只选择与行为策略相似的动作来加以约束[11]。或通过学习当前策略真实值的下界,而无需对行为策略进行限制[12]。一种较为流行的方法是使用生成模型进行估计,该方法在基础实验环境中表现出较为优秀的性能。同时,它可以结合模仿学习中行为克隆的思想,以不同的角度解决离线 RL 分布偏移的问题[13-17]。同样地,基于模型[18-23]的方法在离线 RL 发展中具有重要地位,它能够较大程度上解决分布偏移问题。除此之外,某类特殊的离线 RL 方法也逐渐崭露头角,例如,基于 Transformer 模型[24-28]的离线 RL。随着这些离线 RL 理论的不断发展和完善,离线 RL 技术也在推荐系统[29-31]、导航驾驶[32-34]、自然语言处理[35-36]、机器人控制[37-38]以及医疗[39-40]与能源[41-43]等现实场景中得到了广泛的应用,并被认为是迈向通用人工智能(Artificial General Intelligence,AGI)的重要途径之一。

为了对离线 RL 方法进行全面地分析与总结,从中国计算机学会推荐的国际学术会议与期刊,以及 CNKI 论文数据库中,以“offline reinforcement learning”与“batch reinforcement learning”等关键词进行检索,归纳并整理了这些离线 RL 论文。然后,通过人工审核的方式,对已检索的论文进行筛选,排除与研究问题无关的部分,并将所筛选的论文绘制成图 1。图 1 描述了从 2019 年至 2022 年(截至 2023 年 1 月 1 日),在会议、期刊和网络上,与离线 RL 有关的论文发表数量及刊载情况。其中,论文被收录于 CCF A 类会议 104 篇、CCF B 类会议 14 篇、SCI 一区期刊 5 篇和 SCI 二区期刊 4 篇。根据图 1 可以看出,自从 2020 年离线 RL 理论基础被全面地提出,国内外学者对其关注度日益增加。尤其在 2020 年之后,离线 RL 在学术界引起了广泛的兴趣,其相关论文数量显著增加。

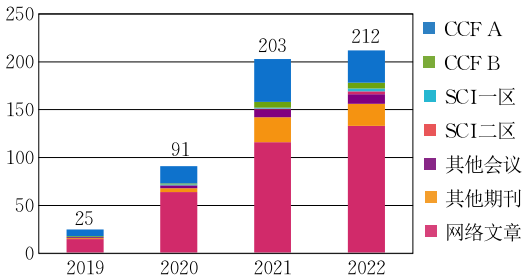


图 1 离线 RL 论文的发表数量及刊载情况

本文梳理了离线 RL 的研究脉络,以其基础理论为指导,重点关注离线 RL 的发展现状与未来趋势。第 1 节为引言;第 2 节介绍离线 RL 的预备知识;第 3 节对离线 RL 进行分类;第 4 节至第 6 节对离线 RL 的核心算法进行对比,描述各类算法的发展历程、研究重点和优缺点;第 7 节对离线 RL 的 3 个实验环境进行解释与比较;第 8 节介绍离线 RL 在现实中的应用;第 9、10 节为离线 RL 的未来展望与总结。

2 预备知识

在本节中,首先介绍了马尔可夫决策过程(Markov Decision Process,MDP)与部分可观测马尔可夫决策过程(Partially Observable Markov Decision Process,POMDP)的数学形式。其次,对部分在线 RL 方法进行梳理。最后,描述了离线 RL 的定义和存在的问题并与在线 RL 进行对比。

2.1 马尔可夫决策过程

RL 作为机器学习的重要组成部分,它以马尔可夫决策过程为理论基础,并以最大化期望回报为目标。用于描述 RL 的马尔可夫决策过程是一个 6 元组 $M=(S,A,P,r,d_0,\gamma)$。其中,S 是一个有限的状态集,A 是一个有限的动作集。$P:S\times A\times S\to[0,1]$是状态转移概率,$P(s_{t+1}\mid s_t,a_t)$表示在状态 s_t 下执行 a_t到达状态 s_{t+1}的概率。$r:S\times A\to\mathbb{R}$ 是即时奖励函数,$r(s_t,a_t)$表示状态 s_t下执行动作 a_t后可以得到的即时奖励。d_0定义了初始状态分布 $d_0(s_0)$。$\gamma\in[0,1]$为折扣因子。MDP 具有马尔可夫性,即 s_{t+1}和 r_{t+1}的每个可能的值出现的概率只取决于前一个

状态 s_t 和动作 a_t，并与历史的状态和动作完全无关。

策略 π 表示状态到动作的映射函数 $\pi: S \rightarrow A$，可分为确定性策略 $a_t=\pi(s_t)$ 和随机性策略 $\pi(a_t|s_t)$。智能体从初始状态分布中产生一个初始状态 s_0，根据策略 π 执行动作 a_0，与环境进行交互获得奖励 r_1，并且转移到下一个状态 s_1，不断重复可以得到一个轨迹长度为 $T+1$ 的状态-动作序列 $\tau=s_0, a_0, s_1, a_1, \cdots, s_T, a_T$。其中，$T$ 可以是无限的。则轨迹分布 $\rho_\pi(\tau)$ 可定义为

$$\rho_\pi(\tau)=d_0(s_0)\prod_{t=0}^{T}\pi(a_t|s_t)p(s_{t+1}|s_t,a_t) \tag{1}$$

带折扣回报定义为从 t 时刻开始到 T 时刻情节结束时的累计奖励：

$$R_t=\sum_{t'=t}^{T}\gamma^{t'-t}r_{t'} \tag{2}$$

其中，γ 存在两种极端情况：如果 $\gamma=0$，则智能体只关注当前状态的奖励。如果 $\gamma=1$，则智能体会更重视未来的奖励。在环境已知的条件下，能够确定所有的未来状态。

RL 的目标在于以最大化期望回报的方式学习并获得最优策略。在状态 s_t 下，智能体遵循策略 π 所得到的期望回报定义为状态值函数（State Value Function）：$V^\pi(s_t)=\mathbb{E}_{\tau\sim\rho_\pi(\tau|s_t)}[R_t|s_t=s]$。状态-动作值函数（State-Action Value Function）也在 RL 中扮演着重要的角色。即在状态 s_t 下执行动作 a_t，智能体遵循策略 π 所得到的期望回报定义为状态-动作值函数：$Q^\pi(s_t,a_t)=\mathbb{E}_{\tau\sim\rho_\pi(\tau|s_t,a_t)}[R_t|s_t=s,a_t=a]$。如果对于所有的状态-动作对，始终存在至少一个策略 π^* 不劣于其他策略，称策略 π^* 为最优策略。尽管最优策略可能不止一个，然而它们共享相同的最优状态-动作值函数：

$$Q^*(s_t,a_t)=\max_\pi\mathbb{E}_{\tau\sim\rho_\pi(\tau|s_t,a_t)}[R_t|s_t=s,a_t=a] \tag{3}$$

最优状态-动作值函数遵循贝尔曼最优方程（Bellman Optimality Equation）：

$$Q^*(s_t,a_t)=r(s_t,a_t)+\gamma\mathbb{E}_{s_{t+1}\sim p(s_{t+1}|s_t,a_t)}[\max_{a_{t+1}}Q^*(s_{t+1},a_{t+1})] \tag{4}$$

当完全可观测状态无法访问时，MDP 可以扩展到部分可观测马尔可夫决策过程[44]。将 POMDP 定义为一个 8 元组 $M=(S,A,\Omega,P,\mathcal{O},r,d_0,\gamma)$。其中，$S,A,P,r,d_0,\gamma$ 与 MDP 中的定义一致。Ω 是观测值空间，在选择某一个动作后，将获得一个系统的观测值 $o\in\Omega$。$\mathcal{O}$ 为观测概率，$\mathcal{O}(a_t,s_{t+1},o_{t+1})$ 表示执行动作 a_t 后，转移到状态 s_{t+1} 时，观测值为 o_{t+1} 的概率。在 POMDP 问题中，智能体在时间步 t，只能获得观测 $o_t=o(s_t)$。

2.2　在线强化学习

近些年，DRL 成为在线 RL 领域的研究热点，Mnih 等人[45]将 Q-Learning 方法与深度学习相结合，提出深度 Q 网络（Deep Q-Network，DQN）。并且 DQN 在 Atari 2600 视频游戏上已经表现出超越人类专家的成果。针对 DQN 只能用于离散动作空间任务的问题，Lillicrap 等人[46]将确定性策略梯度与 DQN 相结合，提出一种适用于大规模连续动作空间的深度确定性策略梯度（Deep Deterministic Policy Gradient，DDPG）算法。为了防止状态-动作值函数被过高地估计，Fujimoto 等人[47]提出了孪生延迟深度确定性策略梯度（Twin Delayed Deep Deterministic Policy Gradient，TD3）算法，在 DDPG 的基础上采用截断的双 Q 学习、延迟的策略更新与目标策略平滑 3 个关键技术。Schulman 等人[48]提出了置信域策略优化（Trust Region Policy Optimization，TRPO）算法，证明了最小化目标损失函数，并且选择合适的步长可以保证策略被单调优化。在此基础上，Schulman 等人[49]进一步提出了近端策略优化（Proximal Policy Optimization，PPO）算法，该方法通过剪枝方式限制新旧策略概率比，避免出现更新步长过大的现象。Haarnoja 等人[50]提出了软性行动者-评论家（Soft Actor-Critic，SAC）算法，并在行动者-评论家（Actor-Critic，AC）算法[51]中引入最大熵，一方面鼓励智能体继续探索，另一方面降低算法对模型与估计误差的敏感性。

2.3　离线强化学习

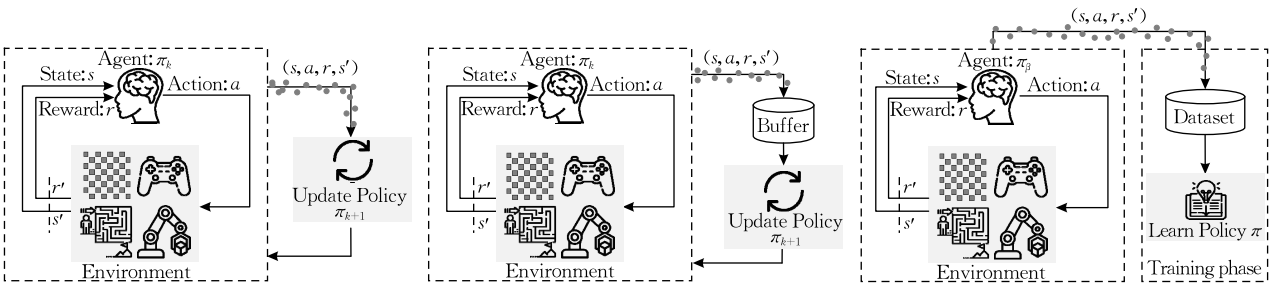
在离线 RL 中，智能体无法在线地与环境进行交互从而收集经验。相对地，智能体从一个静态的数据集 $\mathcal{D}\{(s_t^i,a_t^i,s_{t+1}^i,r_t^i)\}$ 中进行学习。将生成数据集的策略称为行为策略 π_β，并在实际交互时构造一个策略 π 以获得最大的累计奖励。状态 s_t 来自行为策略的状态分布 $s_t\sim d^{\pi_\beta}(s)$，而动作 a_t 根据行为策略 $a_t\sim\pi_\beta(\cdot|s_t)$ 采样获得。下一个状态 s_{t+1} 由状态转移模型 $s_{t+1}\sim P(\cdot|s_t,a_t)$ 决定。$r(s_t,a_t)$ 为状态 s_t 下执行动作 a_t 后得到的奖励。

离线 RL 的目标仍是学习出一个优秀的策略，使其在当前环境中的性能优于离线数据中的行为策略。然而，由于智能体不能在线收集数据，导致出现无法探索的问题。探索与利用是在线 RL 中的经典问题，探索是指智能体在未知环境中探索新的动作和状态，以了解环境并找到更好的策略。而利用是指智能体采取已知的、在之前尝试中表现良好的动作和策略，以最大化当前回报。离线 RL 算法不能进行探索的问题无法得到实质性的解决。继而转变思路，通过最小化贝尔曼均方误差学习 Q 值函数：

$$L(\phi)=\mathbb{E}_{s,a,s'\sim\mathcal{D}}[(Q_\phi(s,a)-(r(s,a)+\gamma\mathbb{E}_{a'\sim\pi}[Q_{\phi'}(s',a')]))^2] \tag{5}$$

当最小化误差时，为了保证 Q 函数在动作 a' 下进行训练，需要当前策略 π 与行为策略 π_β 相同。然而在实践中，要实现这一点是具有挑战性的。因为通常追求找到一种比当前行为策略 π_β 更为优越的新策略 π，从而不可避免地经历分布偏移的过程。此外，数据集质量也十分重要。因为离线 RL 对静态数据集的依赖性，所以要求静态数据集需要具备较高的质量。一个良好的数据集应该具备以下 3 个特点：(1) 多样性。数据集尽量覆盖不同的状态和动作，从而让算法学到更加鲁棒的模型；(2) 准确性。数据集可以准确地反映真实的奖励信号，否则可能导致学习算法产生偏差；(3) 覆盖性。数据集包含多种任务，以保证算法的泛化性与适应性。目前，已有研究者提出 3 类实验环境，这些环境均包含不同类型任务的离线数据集。具体实验环境将在第 7 节展开介绍。

进一步解释在线与离线之间的关系，将在线同策略强化学习、在线异策略强化学习与离线 RL 进行对比，具体如图 2 所示。在图 2 中，(a) 为在线同策略 RL 的学习过程，智能体根据策略与环境进行交互，并产生经验样本完成数据收集。再进行样本学习，不断更新当前的策略 π_k；(b) 为在线异策略 RL 的学习过程，将智能体与环境在线交互产生的数据存储在经验池里，并从经验池中随机采样一些数据用于当前策略 π_k 的更新；(c) 为离线 RL 的学习过程，数据集由未知的行为策略 π_β 产生，并且利用该数据集中的离线数据进行当前策略 π 的学习与更新，从而期望策略 π 通过在线微调等方式，可以在策略部署阶段有不错的表现。



(a) 在线同策略强化学习　(b) 在线异策略强化学习　(c) 离线强化学习

图 2　在线与离线强化学习对比图

3　离线强化学习分类

离线 RL 面临的关键挑战是分布偏移，它被定义为当前策略访问到的状态-动作对与数据集采样得到的状态-动作对分布不一致。具体而言，虽然策略在离线数据的分布下进行训练，但对于部署阶段，训练得到的策略在与环境交互时，访问的状态与之前离线数据相比产生了变化。因此，离线 RL 经常通过最小化公式(5)来学习策略。即使策略能够在训练数据上准确地评估目标，但由于缺失数据或模型偏差，仍可能导致外推误差[11,52](Extrapolation Error)，从而策略引导的状态分布仍可能产生偏离。对于缺失数据而言，如果数据集中缺少某个状态-动作对附近的数据(或者附近的数据量比较少)，那么对于 Q 函数的估计就变得不够准确。若 Q 函数高估了这些训练数据中未曾涉及的状态-动作对，则在实际交互时，当智能体选择最大化 Q 值的动作时，可能选择到实际期望回报非常低的动作。这里称这些数据集分布之外的动作为分布外(Out-Of-Distribution，OOD)动作。对于模型偏差，如果数据集中存储的状态转移不能准确反映真实的 MDP 时，就容易导致误差的产生。

随着研究工作的不断深入，研究者针对分布偏移带来的一系列问题，提出了丰富的离线 RL 方法。根据求解思路的差异，将离线 RL 方法分为无模型(Model-Free)、基于模型(Model-Based)和基于 Transformer 模型 3 大类。此外，离线 RL 还与多智能体[53]、因果推断[54]、元学习[55]以及逆强化学习[56]等方法相结合，但因其发展过程短暂，涉及离线 RL 方面不够深刻，不足以构成一个完整的离线 RL 体系，因此没有对其他支线进行讨论。

具体而言，这些方法有着不完全相同的关注点和待解决的问题，实现了在解决分布偏移问题上的逐步改进。(1) 无模型离线 RL 方法。在面对分布偏移问题时，该方法通常设计一种受离线数据限制或惩罚的算法，而无法使用额外数据。因为数据集本身总是限制学习策略进行适当泛化，所以无模型离线 RL 方法学到策略的一般较为保守。然而，由于离线数据通常非常有限，学习模型被认为是不稳定的，所

以大多数方法都采取保守策略；(2) 基于模型的离线 RL 方法。该方法能够利用环境模型来预测不同状态下的奖励与转移概率[57]。虽然状态转移模型可以用于额外的交互，来产生额外的训练数据来提高策略性能。但是该模型是基于行为策略的，所以仍存在分布偏移问题。因此在某种意义上，该方法利用了模型的泛化能力来执行一定程度的探索，进一步解决了分布偏移问题；(3) 基于 Transformer 模型的离线 RL 方法。该方法同时对环境和行为策略进行建模，无需考虑无模型中的限制条件。这有效地避免了外推误差问题，使得离线 RL 的分布偏移问题得到实质性的解决。同时，方法关注 RL 中的经典问题，基于贝尔曼最优方程的更新方式传播得较慢，并且极其容易受到扰动的影响。在长时序与稀疏奖励环境中效果不佳。因为 Transformer 本身具有较强的表征能力，拟合数据后存在一定的泛化性，所以更容易在长时序与稀疏奖励环境中取得良好的性能。从而实现分布偏移问题由无模型向基于模型和 Transformer 模型的递进求解。

下面将用 3 节内容分别阐述无模型、基于模型与基于 Transformer 模型的离线 RL 方法各自特点与研究内容。为了直观地展现离线 RL 方法的分类情况，绘制分类图，具体如图 3 所示。离线 RL 方法依据求解思路不同划分类别，可以更加清晰地理解这些方法的区别。此外，具体的细分类别将在相应章节中进行详细介绍。

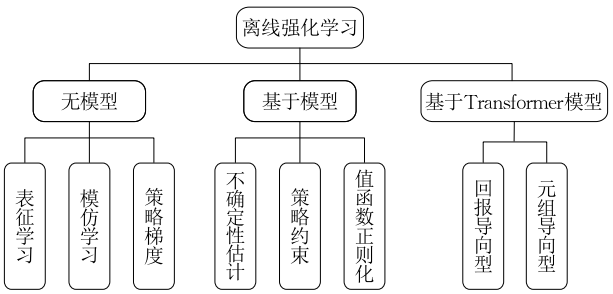
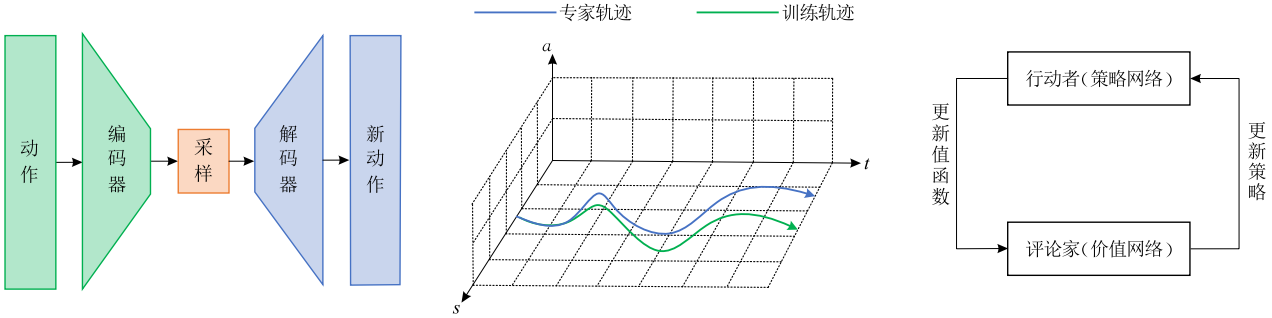


图 3　离线强化学习分类

4　无模型

在传统无模型 RL 方法中，状态转移概率与智能体所处的环境模型是未知的。这就需要智能体与环境进行交互，采集更多的轨迹数据来改进策略。而无模型离线 RL 方法无法与环境交互，虽然可以直接从静态数据集中学习策略，但是可能造成分布偏移的问题。因此，无模型离线 RL 的关键在于如何解决分布偏移问题，使得离线数据的分布和目标策略相应的数据分布尽可能一致。为避免分布偏移引起的外推误差，离线 RL 方法通过对学习策略施加惩罚，受行为策略的影响，目标策略只能在有限的静态数据集中学习。因此，策略在静态数据集外的泛化能力较弱，且学习所得策略通常较为保守。

根据近几年无模型离线 RL 方法的技术发展路线，从不同方法原理出发，将其细分为表征学习 (Representational Learning)、模仿学习 (Imitation Learning) 和策略梯度 (Policy Gradient) 这 3 个子类别。表征学习将离线数据映射到一个更具信息含量和抽象性的表示空间，以提高对策略的理解和泛化[58]。而模仿学习根据行为策略中的样本来学习策略。策略梯度则是通过参数化策略，使用函数近似对策略进行拟合。为了直观地对比 3 种无模型离线 RL 方法的差异，绘制图 4 如下。其中，(a) 为表征学习中的一种动作表征结构，通过该结构生成与数据集相似的新动作集合；(b) 为模仿学习中的行为克隆方法，使训练轨迹接近专家轨迹；(c) 为策略梯度中经典的行动者-评论家 (AC) 框架。



(a) 动作表征结构　　(b) 模仿学习　　(c) 基于AC框架的策略梯度

图 4　无模型离线 RL 的 3 种方式

4.1　表征学习

表征学习 (表示学习) 作为一种机器学习方法，其目标是学习数据的内在特征，寻找一种有效的数据表示。尽管离线 RL 和表征学习是两种不同的学习范式，但在面对复杂和大规模问题时，通过表征学习可以显著提高离线 RL 过程的计算效率。

此外,在处理非独立同分布的数据时,表征学习表现出良好的性能,能够有效地解决在函数近似过程中的累积误差问题。

4.1.1 动作表征

动作表征的目标是将原始的高维动作空间映射到低维表征空间,从而降低动作空间的复杂性。基于动作表征的离线 RL 方法通常采用编码器-解码器结构,使其更好地捕捉动作的内在特征。

Fujimoto 等人[11]受到变分自动编码器[59](Variational Auto-Encoder,VAE)架构的启发,将生成模型与离线 RL 相结合,并提出一种批量约束 Q 学习(Batch Constraint Q-Learning,BCQ)算法,该算法可以直接对策略进行限制。具体来说,通过类似于行为策略 π_β 选择的动作来约束当前的学习策略 π:

$$\pi(s)=\underset{a_i+\xi_\theta(s,a_i,\Phi)}{\arg\max}\ Q_\phi(s,a_i+\xi_\theta(s,a_i,\Phi)) \quad (6)$$

其中,$a_i\sim G_\omega(s)$,$i=1,2,\cdots,n$。$G_\omega(s)$为估计 π_β 的参数化生成模型,$\xi_\theta(s,a_i,\Phi)$为扰动模型。因其依赖于行为策略 π_β 估计,可以看作是一种显式约束。该算法作为离线 RL 领域的开创性工作之一,不仅提出了外推误差的概念,还提供了解决离线 RL 问题的初步方案。从实验结果来看,在 Gym-MuJoCo 连续环境下,BCQ 算法的学习性能比其他离线算法性能更好,但是只与在线无决策扰动的 DQN 性能相当。随后,完善了在 ALE 环境下的离散 BCQ[52]算法。

虽然 BCQ 使用的 VAE 与扰动模型解决了外推误差的问题,但对于一些 OOD 的状态-动作对无法很好地拟合。Kumar 等人[60]分析其原因,指出基于贝尔曼方程的迭代能够产生自举误差(Bootstrapping Error),并提出一种减少自举误差积累(Bootstrapping Error Accumulation Reduction,BEAR)算法来保证当前的学习策略与静态数据集中的分布相匹配。在该算法中延续 VAE 模型的使用,并引入支持匹配(Support Matching)的概念,将策略动作限制在训练分布的支持集中。对比之前的 BCQ 算法,可将 BCQ 看作是一种分布匹配(Distribution Matching)的方式。分布匹配约束学习策略的分布与行为策略相匹配,而支持匹配约束是将学习策略选择的动作限制在由行为策略选择的动作的支持范围内。此外,BEAR 使用一种基于采样的最大均值差异(MMD)散度,该散度作为距离度量来约束策略。在 BEAR 提出的数据集中,该算法表现出良好的性能。但是在后续统一数据集的 D4RL 中,BEAR 算法的表现不及 BCQ 算法。探究其原因可知,BEAR 算法的实验参数过多,整体框架相对复杂,这导致其性能的不稳定。

由于将策略限制在数据支持集内的 BEAR 算法未能有效地避免 OOD 动作,Wu 等人[61]提出了支持策略优化(Supported Policy Optimization,SPOT)算法。该算法采用条件变分自动编码器[62](Conditional Variational Auto-Encoder,CVAE)来明确估计正则化项中的行为密度,达到直接计算散度的目标。此外,SPOT 的正则化项可以嵌入到任何异策略 RL 算法中。然而,在 Gym-MuJoCo 和 AntMaze 任务中,SPOT 未能显著提高任务的性能。

不同于以上算法,Zhou 等人[63]提出潜在动作空间策略(Policy in the Latent Action Space,PLAS)算法。该算法使用 CVAE 对行为策略进行重建。通过使用 CVAE 将状态映射到潜在行为空间中,从而使得学习到的行为策略更加鲁棒和通用。在这个过程中,CVAE 不需要使用任何关于行为策略的标签信息,而是通过最大化重构概率和 KL 散度来学习潜在变量表示。同时,延续 BEAR 的思想,将策略限制在数据集的支持范围内。与 BEAR 算法不同,该算法采用一种隐式约束,并通过潜在动作空间训练策略。这种方法的优点在于,它可以在构建潜在空间的过程中自然满足,并且不影响算法其他部分的优化,同时也不受行为策略分布的限制。实验结果表明,在大部分 Gym-MuJoCo 任务下,PLAS 算法优于 BCQ 和 BEAR 算法的性能。此外,将该算法应用在真实机器人中,可以为未来机器人的技术发展带来巨大的推动作用。

对于使用 MMD 散度约束策略的 BEAR 而言,该算法在连续控制任务上表现出次优性能。基于此,Chen 等人[64]在 PLAS 基础上提出潜在空间优势加权策略优化(Latent-Space Advantage-Weighted Policy Optimization,LAPO)算法。该算法学习了潜在变量策略,使得该策略能够从潜在空间的先验分布中采样,以生成高优势动作。此外,LAPO 预先训练了一个 VAE 的模型,以重构生成多模态数据的行为策略。在实验部分,该算法不仅与基线算法在 D4RL 中进行了性能比较,还在多模态条件下进行了验证,为多模态离线 RL 提供了一种新的思路。

为了更好地学习大规模离散动作空间的离线 RL 任务,Gu 等人[65]提出了一种动作的行为度量(Behavioral Metric of Actions,BMA)框架。该框架测量了动作之间的行为关系和数据分布关系,旨在

明确量化行为对环境影响的相似程度。通过动作编码器，以自监督的方式学习动作表示，学习到的动作表示可以与任何离线 RL 算法相结合。在实验部分，该框架不仅在迷宫世界的任务中提高了性能，还验证了在现实应用中的有效性。

在离线 RL 中，保守地估计值函数是必要的。但是过于保守的策略极大程度地限制了值函数的泛化性。因此，为了提高离线RL中值函数的泛化能力，Lyu等人[66]提出轻度保守 Q 学习（Mildly Conservative Q-Learning，MCQ）算法。该算法通过给 OOD 动作赋予合适的伪 Q 值来进行训练。该伪 Q 值使用最大 Q 值减去一个正值，以确保在足够保守性的前提下提升泛化能力。尽管该算法利用了 CVAE 来对行为策略进行建模，仍然可能导致 OOD 动作的问题。然而，理论上证明了 MCQ 相较于行为克隆在性能上更具优势，并且不容易高估 OOD 动作。

Lou 等人[67]设计了一种基于互信息的离线 RL（Offline RL Mutual Information，ORL-MI）算法，从一个新的角度出发来提高离线 RL 中的泛化性。该算法采用动作嵌入表征模型对动作进行编码和解码，从理论层面探讨了值函数在作用空间上泛化能力的提高，并提供了一个基于信息论的解释。在实验部分，ORL-MI 在 Gym-MuJoCo 任务中的有效性得到验证，但是实验结果并未清晰展示动作是如何进行泛化的。

先前基于编码器-解码器架构的离线表征学习大多采用了 VAE 及其衍生模型。这类模型使用似然函数的变分下界代替真实的数据分布，因而只能得到真实数据的近似分布。相比之下，流模型[68]能够明确地学习数据分布，并且训练过程较为稳定。基于此，Akimov 等人[69]利用保守标准化流（Conservative Normalizing Flow，CNF）算法，用于构建对离线 RL 有用的潜在动作空间。通过在网络模型的最后一层添加可逆激活函数 Tanh，使潜在策略可以充分利用整个潜在空间。

类似地，Wang 等人[70]创新地将扩散模型引入了离线RL中，提出扩散Q学习（Diffusion Q-Learning，Diffusion-QL）算法，该算法利用条件生成模型来表示策略。具体来说，使用基于多层感知器（MLP）的去噪扩散概率模型（DDPM）[71]作为策略。一方面鼓励扩散模型对与训练集中分布相同的动作进行采样，另一方面对 Q 值较高的动作进行采样。实验结果表明，在 D4RL 实验环境中，Diffusion-QL 算法可以获得更好的性能。同时，该算法优于其他基于 VAE 与 CVAE 等生成模型的离线 RL 算法。此外，还进行了多模态实验，证明了扩散模型对于多模态任务的重要性。该算法的不仅可以处理高维、连续的动作空间，而且对于部分可观测的状态也有较好的性能。但是可能需要更多的实验来进一步验证其优越性。

近些年，安全强化学习[72]（Safe Reinforcement Learning，SRL）领域取得了显著的进展。Dong 等人[73]引入 SRL 的概念，并用生成对抗网络（Generative Adversarial Network，GAN）替代 BCQ 算法中的 VAE 生成模型，提出了 SGBCQ 算法。该算法将安全限制条件融入离线 RL 中，确保了智能体的安全性，从而避免因环境不确定性导致的意外行为。同时，选择不同的基于 GAN 的衍生方法对性能产生显著的影响。但是由于 GAN 极容易引起模式塌缩问题，应该进一步对该问题提出解决方法。此外，该算法在 NeoRL 环境中进行实验，证明了其优越性。

4.1.2　状态表征

状态表征的目标是学习一种表示方法，以捕捉环境中的关键信息，同时减少冗余和噪声，使得智能体能够更好地理解和应对不同的环境。

为了提高值函数网络的泛化性，Weissenbacher 等人[74]提出库普曼前向 Q 学习（Koopman Forward Conservative Q-learning，KFC）算法。该算法利用环境动力学中的对称性来引导数据增强策略，并在训练期间对静态数据集进行扩展。KFC 基于动态控制系统对称性和数据对称位移的理论结果，因此适用于具有可微状态转换的系统和具有库普曼算子的双线性化模型的系统。然而，这也带来了一些局限性，导致 KFC 无法应对不连续的任务。在实验部分，KFC 不仅在 D4RL 环境中进行了全面的验证，还在 MetaWorld[75]和 RoboSuite[76]环境下进行了测试。

由于静态数据集大小和质量的限制，提高从高维观测的离线数据中的零样本泛化性能是一个具有挑战性的问题。为应对这一挑战，Mazoure 等人[77]提出了一种广义相似性函数（Generalized Similarity Functions，GSF）算法。该算法利用对比学习来训练策略，以捕捉给定状态下任意瞬时累积量的未来行为。选择瞬时累积量的方式决定了行为相似性的性质。为了评估离线 RL 算法在零样本泛化方面的性能，Mazoure 等人设计了离线版本的 Procgen[78]环境。通过实验证实，GSF 在处理具有挑战性的基于像素的控制任务时表现出有效性。

离线 RL 算法不仅难以处理有限的高维数据，

例如具有连续动作空间的视觉问题，还由于高度重复使用数据，进一步加剧了价值网络的隐式参数化不足问题。为了应对这些问题，Zang 等人[79]提出一种行为先验表征(Behavior Prior Representation，BPR)算法。该算法利用行为策略来学习状态表示，而不是指定特定的属性。通过在数据集中模仿动作来学习状态表示，并在固定表示上使用任何现有的离线 RL 算法来训练策略。尽管 BPR 算法在学习对视觉干扰很强的表示方面表现出了有效性，但当训练环境与评估环境存在显著差异时，其泛化性仍然需要进一步提高。

在基于表征学习的离线 RL 中，常常使用生成模型来学习状态的潜在表示。通过在潜在空间中操作，这些模型能够生成新的状态和动作。目前，应用于离线强化学习的生成模型包括 VAE、GAN、流模型、扩散模型以及它们对应的变种模型。这些表征学习方法有助于智能体学习环境中的结构和规律，从而提高策略的鲁棒性。

4.2 模仿学习

模仿学习是指从专家演示的数据中学习，通过学习专家的决策过程，学习到接近于专家的策略。这使得利用专家数据的模仿学习相比 RL 具有更高的样本效率。模仿学习通常有两种类型，一种是直接从专家演示的数据中学习策略的行为克隆[80-81](Behavior Cloning，BC)方法。另一种通过专家示范样例重构奖励函数的逆强化学习[82-83](Inverse Reinforcement Learning，IRL)方法。

目前，一种常用的方法是将 BC 思想引入离线 RL 学习中，其核心内容是模仿行为策略 π_β。BC 方法可以精确地复制这些策略，并且基于最小化学习策略与行为策略的距离度量来逼近专家策略。即 $\min D(\pi_\beta(\cdot|s),\pi_\theta(\cdot|s))$。其中 D 为距离度量。但是，在离线 RL 中直接访问专家行为策略是很困难的，因为已有的离线轨迹数据集中可能包含不良行为。为了处理不良行为，基于模仿学习的离线 RL 方法分为直接过滤法与条件策略模型。

4.2.1 直接过滤法

直接过滤法是指直接从数据集中滤除不良行为，只模仿优质行为。为了解决这个问题，通常使用值函数和启发式方法来筛选数据集中的不良行为。例如，利用值函数评估每个轨迹的质量，并且选择高质量的轨迹进行训练。启发式方法则利用专家演示来指导智能体在数据集中选择最有价值的行为。

Chen 等人[13]提出最佳动作模仿学习(Best Action Imitation Learning，BAIL)算法。为了简化和提高性能，该算法通过学习状态值函数 $V(s)$ 来选择性能高的动作，而不涉及在状态-动作空间上最大化 Q 函数。通过将 $V(s)$ 拟合到数据集的近似上包络(Upper Envelope)，可以近似出满足贝尔曼方程的最优值函数。再通过过滤不良行为的 BC 方法来训练策略网络。实验结果表明，BAIL 算法在 Gym-MuJoCo 任务下的性能较 BCQ[11] 和 BEAR[60] 算法均有提升。但是，由于该算法提出的奖励回报修正方法具有局限性，在 OpenAI 的迷宫世界或者 ALE 环境中存在一些限制。

Siegel 等人[14]将优势加权行为模型(Advantage-Weighted Behavior Model，ABM)作为最大化后验策略优化[84](Maximum Posterior Policy Optimization，MPO)的先验方法，提出了 ABM-MPO 算法。该算法遵循策略迭代的流程，在策略评估阶段，对状态-动作值函数进行评估。在策略改进阶段，优化学习策略保证它与离线数据分布尽可能接近。当离线数据是来自多种任务、多种非完美策略时，可以针对不同的轨迹使用不同权重的优势函数，从而筛选掉导致学习策略表现变差的轨迹。优势函数和目标函数分别为

$$A^\pi(s,a)=\sum_{j=t}^{N-1}\gamma^{j-t}r(s_j,a_j)+\gamma^{N-t}\hat{V}^{\pi_i}(s_N)-\hat{V}^{\pi_i}(s) \tag{7}$$

$$J(\theta_{\text{abm}})=\mathbb{E}_{s,a\sim\mathcal{D}}\Big[\sum_{t=1}^{|\tau|}\log\pi_{\theta_{\text{abm}}}(a|s)f(A^\pi(s,a))\Big] \tag{8}$$

其中，f 是一个非负的递增函数，它用于赋予不同优势函数以不同的权重，在这里 $f=1[A^\pi(s,a)>0]$。在学习 ABM 先验之后，使用 MPO 来改进学习策略 π_θ，通过 KL 散度约束到 $\pi_{\theta_{\text{abm}}}$。实验结果表明，在 Gym-MuJoCo 任务中，ABM-MPO 算法可以提升大部分次优数据的学习性能，并且与基线相比其性能有所提升。此外，通过模拟一个 Sawyer 机械臂，实现了在多任务环境中验证算法性能的目标。类似地，Wang 等人[15]提出一种评论家正则化回归(Critic Regularized Regression，CRR)算法。该算法仍然使用优势函数的估计值在数据集中为 BC 选择最佳动作，类似式(9)。不同的是，CRR 算法选择了 $f=\exp(A^\pi(s,a)/\beta)$ 和更加悲观的优势函数。CRR 算法不依赖观察到的奖励来进行优势估计，而是引入了一种评论家加权策略(CWP)技术，并使用学习到的评论对算法进行改进。该算法在 RL Unplugged 实验环境下进行实验，结果表明，它能够扩展到具有高维状态-动作空间的任务，并且有效处理低质量数据集。

4.2.2　条件策略模型

基于条件策略模型法是一种不需要专家行为的方法，可以完全离线地学习策略。这种方法的基本思想是，将策略建模为一个条件概率分布，具体可表示为 $\pi_\theta(a_t|s_t,\omega)$。其中，$\omega$ 以剩余轨迹为条件，即 $\omega\sim f(\cdot|\tau_{t:H})$。通过 BC 学习条件行为策略模型，则目标函数可表示为

$$J(\theta)=\mathbb{E}_{\tau\sim p_{\pi_\beta}(\cdot),t\sim\mathrm{Unif}(1,H),\omega\sim f(\cdot|\tau_{t:H})}[\log\pi_\theta(a_t|s_t,\omega)] \tag{9}$$

其中，Unif(1，H)为均匀分布。如何定义适当的结果函数 f 是一个关键问题。在实践中，需要根据任务的特定需求和数据集的特征选择合适的函数。

不同于以上直接过滤不良行为的方法，Emmons 等人[16]采用条件策略模型来学习策略，提出了一种通过监督学习实现的(RL via Supervised learning，RvS)算法。该算法采用两种不同的条件策略模型，一是目标条件策略，其结果函数为

$$f(\omega|\tau_{t:H})=\mathrm{Unif}(s_{t+1},s_{t+2},\cdots,s_H) \tag{10}$$

二是奖励条件策略，其结果函数为

$$f(\omega|\tau_{t:H})=1\left[\omega=\frac{1}{H-t+1}\sum_{t'=t}^{H}r(s_{t'},a_{t'})\right] \tag{11}$$

实验结果表明，目标条件策略可以较好地实现次优轨迹拼接，因此在 AntMaze 与 FrankaKitchen 任务下的效果较好。然而，奖励条件策略在这些复杂环境中的表现非常差，但在 Gym-MuJoCo 任务下表现良好。此外，该算法只需简单地使用两层 MLP 网络与正则化技术，就可以在性能上与具有复杂模型的算法相媲美。然而，RvS 算法也存在一定的局限性，在 Gym-MuJoCo 任务下，针对随机数据的算法结果并不理想。

为了使算法更具稳定性和泛化性，Xu 等人[17]提出一种策略引导的离线 RL(Policy-guided Offline RL，POR)算法。该算法将传统离线 RL 中的最大化奖励策略解耦成一个指导策略和一个执行策略。在训练过程中，以监督和解耦的方式，仅使用静态数据集中的数据来学习指导策略和执行策略。在评估过程中，指导策略通过指导执行策略来选择动作，从而最大化奖励。根据实验结果来看，POR 在 Gym-MuJoCo 和 AntMaze 任务下的性能提升并不明显。但面对新任务时，通过解耦的训练方式可以复用执行策略，以更少的计算资源完成任务的迁移。

BC 方法作为基础的模仿学习算法，不仅使用简单，并且易于实现。因其无需设计复杂的奖励函数，这使得该方法在一些任务上表现良好，特别是在可以获得大量优秀历史决策数据的条件下[85]。大多数基于模仿学习的离线 RL 问题本质上是通过 BC 方法实现的。该方法在学习过程中扮演了重要的角色，但其必须从专家演示的数据中进行学习。基于此，离线 RL 通常采用直接过滤掉次优动作的方法，并利用传统的监督回归方法计算损失。因此，Kumar 等人[86]对于在提供专家或接近专家的数据情况下，离线 RL 方法是否比 BC 方法更可取提出了疑问。实验结果表明，在不同类型的专家策略下，离线 RL 方法在各种实际问题领域上都优于 BC。尤其在稀疏奖励或有噪声的次优专家数据上，离线 RL 方法显著优于 BC。

在基于模仿学习的离线 RL 中，通常使用高质量轨迹进行训练，可有效避免训练过程中的不稳定性和收敛性问题，从而提高学习效率和训练的准确性。但是该方法需要妥善处理不良行为轨迹，如果无法获得足够高质量的数据，模型的性能将受到限制。

4.3　策略梯度

策略梯度是一种经典的策略提升方法，它通过持续计算策略参数的梯度，以使策略期望总奖赏最大化，并相应地更新策略参数，直至收敛于最优策略[87]。与随机集成混合(Random Ensemble Mixture，REM)[88]等基于值函数的算法相比，基于策略梯度的离线 RL 方法适用范围更广，具有更好的收敛性。然而，在许多复杂的现实场景中，策略梯度易收敛到局部最优解。针对此问题，DRL 算法通常以经典 RL 方法中的 AC 框架为基础进行拓展。同样地，在基于策略梯度的离线 RL 算法中，AC 框架是一个常用的底层算法。该框架可分为基于行动者的策略网络与基于评论家的价值网络。并且通过不断地更新值函数和策略函数，来寻找最优策略。大部分的算法只在一个网络上进行改进，只有少数算法在两个网络上都进行了改进。因此，根据在哪个网络上进行主要改进来划分基于策略梯度的离线 RL 方法。

4.3.1　策略网络

对策略进行正则化并在基础 AC 框架上改进策略网络，可以使得算法更加稳定和鲁棒，从而在实际应用中发挥更大的作用。

Wu 等人[89]提出行为正则化行动者-评论家方法(Behavior Regularized Actor-Critic，BRAC)。该算法指出策略进行正则化的方法主要有两种：一是在函数中加入惩罚项(Value Penalty，VP)；二是在策略中加入正则项(Policy Regularization，PR)。此外，归纳出 4 种适合的距离度量分别为 MMD、KL、

f、Wasserstein[90]，并进行了全面的超参数敏感分析。实验结果表明，BRAC-VP 在大多数情况下的性能略优于 BRAC-PR。而 4 种距离度量整体差别不大，MMD 与 KL 只在 Hopper-v2 任务中有较明显的优势。BRAC 更多的是对先前内容的总结与对比。

为了降低离线 RL 算法的复杂性，Fujimoto 等人[91]进一步提出了 TD3＋BC 算法。该算法在 TD3 的基础上引入了一个行为克隆正则项 $\pi(s)-a$，类似于一种距离约束，以增强算法的性能。在实践中，对其进行状态归一化。TD3＋BC 算法不仅结构简单，而且相比其他离线 RL 方法，它训练时间较短，从而极大地节约了时间成本。

在 TD3＋BC算法基础上，Ran 等人[92]提出了基于数据集约束的策略正则化(Policy Regularization with Dataset Constraint，PRDC)算法。在更新特定状态的策略时，该算法在离线数据集中搜索与当前状态-动作对最接近的样本，并利用该样本的动作限制策略。即最小化状态-动作对和数据集之间点到集的距离：

$$L_{\mathrm{DC}}(\theta)=\mathbb{E}_{s\sim\mathcal{D}}[d^{\beta}_{\mathcal{D}}(s,\pi_\theta(s))]$$
$$=\mathbb{E}_{s\sim\mathcal{D}}[\min_{(\hat{s},\hat{a})\in\mathcal{D}}\|(\beta s)\oplus\pi_\theta(s)-(\beta\hat{s})\oplus\hat{a}\|] \quad (12)$$

其中，$\oplus$表示向量拼接。β是用来平衡状态和动作差异的超参数。此外，该算法使用 KD-Tree 方法显著减少了计算成本。实验结果表明，PRDC 算法在 Gym-MuJoCo 和 AntMaze 任务中表现优异。

为了对 RWR[93]算法进行改进，Peng 等人[94]提出了一种隐式约束的优势加权回归(Advantage Weighted Regression，AWR)算法，其核心思想是将策略优化过程看成是极大似然估计问题。该方法通过对要学习的策略施加惩罚，得到目标函数为

$$J(\theta)=\mathbb{E}_{s\sim d^{\pi_\beta}(s)}[\mathbb{E}_{a\sim\pi_\theta(a|s)}[\hat{A}^{\pi}(s,a)]+\eta(\varepsilon-D_{\mathrm{KL}}(\pi_\theta(\cdot|s),\pi_\beta(\cdot|s)))] \quad (13)$$

上述函数对 π_θ求偏导，最终得到最优策略 π^*：

$$\pi^*(a|s)=\frac{1}{Z(s)}\pi_\beta(a|s)\exp\left(\frac{1}{\eta}\hat{A}^{\pi}(s,a)\right) \quad (14)$$

尤其在策略提升阶段，采用了优势函数 $\hat{A}^{\pi}(s,a)$代替了状态-动作值函数 $\hat{Q}^{\pi}(s,a)$进行似然估计。与 AWR 算法类似，Nair 等人[95]提出优势加权行动者-评论家(Advantage Weighted Actor-Critic，AWAC)算法。该算法与 AWR 之间的关键区别在于价值估计不同。AWR 通过蒙特卡洛或时序差分来估计行为策略的状态值函数，AWAC 则是通过自举来估计当前策略的状态-动作值函数。虽然整体改动不大，但在实验的所有任务中，AWAC 算法的样本效率明显高于 AWR。此外，该算法还提供了一种特别的在线微调场景，从而形成离线预训练在线调整的模式。

在正则化行为值估计[96](Regularized Behavior Value Estimation，R-BVE)的基础上，Brandfonbrener 等人[97]创新性地提出单步(One-Step)算法。通过单步方法，解决了离线 RL 中多步(Multi-Step)方法遇到的迭代误差利用(Iterative Error Exploitation)等问题。对单步方法只进行一次行为策略的评估，再进行一次策略改进。然而，对于多步迭代方法，通过不断地迭代直到最优。BCQ 和 CQL 等算法都是通过多步迭代，交替地进行策略评估与策略改进过程。对比两种方法，单步方法依赖于单步策略迭代。具体而言，该方法涉及对行为策略的价值函数进行拟合，随后采用相应的贪心策略。另一方面，单步方法也可以完全避免使用价值函数，而是采用行为克隆的目标。该方法通过学习行为策略的价值函数或不使用价值函数，避免了查询分布外的动作。因此，在处理计算资源有限的场景中，单步方法更为适用。然而，多步方法则通过多次迭代执行真正的动态规划，当环境提供的数据覆盖率较高时，原则上能够更有效地学习最优策略。因此，在需要捕捉长期依赖关系的场景中，多步方法更为适用。

4.3.2 价值网络

同样地，正则化等技术也可以应用在值函数估计上，对价值函数进行估计也在一定程度上帮助策略网络更好地更新，与其形成相辅相成的关系。

为了进一步扩展研究，Kumar 等人[12]将重点聚焦于值函数上，提出了保守 Q 学习(Conservative Q-Learning，CQL)算法。该算法关键在于避免由于分布偏移导致的 Q 值过高估计问题。故 CQL 算法直接从值函数出发，在 Q 值的基础上添加正则化项，旨在得到真实状态-动作值函数的下界估计。为了获得该下界，应添加正则项：

$$\max_{\pi_\mu}\alpha(\mathbb{E}_{s\sim\mathcal{D},a\sim\pi_\mu(a|s)}[Q(s,a)]-\mathbb{E}_{s\sim\mathcal{D},a\sim\hat{\pi}_\beta(a|s)}[Q(s,a)])+\mathcal{R}(\pi_\mu) \quad (15)$$

CQL 算法使用两个目标之和训练 Q 函数：正则化项与时序差分(TD)误差。正则化项在执行 OOD 动作时最小化 Q 值，同时最大化行为策略 π_β下的 Q 值，故得到损失函数为

$$L(Q)=\min_{Q}\max_{\pi_\mu}\alpha(\mathbb{E}_{s\sim\mathcal{D},a\sim\pi_\mu(a|s)}[Q(s,a)]-\mathbb{E}_{s\sim\mathcal{D},a\sim\hat{\pi}_\beta(a|s)}[Q(s,a)])+\frac{1}{2}\mathbb{E}_{s,a,s'\sim\mathcal{D}}[(r(s,a)+\gamma\mathbb{E}_{\pi}[\bar{Q}(s',a')]-Q(s,a))^2]+\mathcal{R}(\pi_\mu) \quad (16)$$

其中，$\hat{\pi}_\beta(a|s)$是行为策略 π_β的估计，π_μ与当前学习的策略相同，而此时策略的值就是真实值函数的下界。类似在线 RL 中策略迭代的过程，直接将 π_μ定义为能够最大化 Q 值的策略。$\mathcal{R}(\pi_\mu)$为策略 $\pi_\mu(a|s)$的正则化项。因而，进一步提出了基于 AC 框架的 CQL 算法，在经典 CQL 算法上增加对学习策略的训练，并对应修改 Q 函数的目标。因此，CQL 可以看作是策略梯度法。实验结果表明，CQL 算法在 D4RL 实验环境和 ALE 环境中都取得了良好的性能。特别地，在 D4RL 中的 AntMaze 和 FrankaKitchen 任务下，该算法显著优于基于表征离线 RL 方法中的 BCQ[11] 与 BEAR[60] 算法。说明 CQL 算法更加适用于复杂任务、稀疏奖励和多模态数据分布的情况。该算法的提出对于推进离线 RL 方法在连续控制任务中的发展具有重要意义，并成为研究的热点之一。

Kostrikov 等人[98] 提出一种基于 Fisher 信息距离[99] 的行为正则评论家(Fisher-Behavior Regularized Critic，Fisher-BRC)算法。该算法通过行为正则化的方式让所学习的策略尽量与行为策略相似。将价值网络参数化为生成离线数据的对数行为策略，并增加一个额外的状态-动作偏移项。此外，使用 Fisher 信息距离来约束策略。该距离通过采样分布而不是积分来衡量两个分布的差距，有效地减少了计算量。对数和表达式项对于连续动作来说是难以处理的，在 CQL[12] 算法中，通过具有重要性权重的蒙特卡洛采样来解决该问题，其中当前的训练策略被用来抽取样本。但是这个过程容易给行动者-评论家的训练增加计算负担。相比之下，Fisher-BRC 使用一种新颖的评论家表述和 Fisher 信息距离，从而避免了繁琐的数值积分计算。由实验结果可知，正则化系数 λ 对算法性能至关重要，λ 越大，学习策略越接近行为策略。但是，如果没有正则化 $\lambda=0$，Fisher-BRC 算法可能在大多数任务中崩溃。当正则化系数太高即 $\lambda=1$，学习策略将受到过度约束。

在 One-step[97] 的思想基础上，通过避免学习 OOD 的动作，而是利用已知的状态-动作对进行学习，并使用 SARSA 的方式重构策略和值函数。由此，Kostrikov 等人[100] 提出了隐式 Q 学习(Implicit Q-Learning，IQL)算法。该算法引入一个单独的状态值函数，并采用预期回归(Expectile Regression)损失。在策略评估阶段，没有使用行为策略采样得到的动作来更新 Q 函数，而是采用状态值函数逼近器作为目标。在策略改进阶段，利用 AWR[94] 中指数优势权重策略抽取(Policy Extraction)的算法。从实验结果看，IQL 算法在 D4RL 环境中具有良好的性能，尤其在该环境下的 AntMaze 和 Adroit 任务中表现出显著的优势和可靠的性能。IQL 算法不仅具有 One-step 算法中无需对 OOD 动作进行任何查询的特点，而且还可以执行多步动态规划(Multi-Step Dynamic Programming)。该算法易于实现，且计算效率高，只需要为评论家增加一个额外的正则项。

在 BEAR 算法基础上，Wu 等人[101] 提出一种不确定性权重行动者评论家(Uncertainty Weighted Actor-Critic，UWAC)算法。该算法通过蒙特卡洛丢弃(Monte Carlo Dropout，MC dropout)来估计状态-动作对的不确定性，进而在行动者和评论家网络的损失中，利用其特性对 Q 估计值加权来降低 OOD 样本的权重。基于这一思想，模型不确定性可以通过 Q 值估计的方差近似得到：

$$\begin{aligned} U(s,a)&=\mathrm{Var}[Q(s,a)]\\ &\approx\sigma^2+\frac{1}{T}\sum_{i=1}^{T}\hat{Q}_t(s,a)^{\mathrm{T}}\hat{Q}_t(s,a)-\\ &\quad\mathbb{E}[\hat{Q}_t(s,a)]^{\mathrm{T}}\mathbb{E}[\hat{Q}_t(s,a)] \end{aligned}\tag{17}$$

其中，T 为采样次数，第 1 项 σ^2是固有噪声，即数据的不确定性。第 2 项是模型关于预测的不确定性，第 3 项是预测均值，故第 3 项减去第 2 项就是模型关于 OOD 状态-动作对的不确定性。同时，对新策略 π' 加权得到 $\pi'(a|s)=\beta\pi(a|s)/\mathrm{Var}[Q_0^{\pi'}(s,a)]Z(s)$。其中，$Z(s)$是归一化因子，当不确定性越大时，权重越小，则考虑使用此策略的机会越小，选择对应动作的概率也越低。此外，对 AC 算法中的贝尔曼更新方式进行了修改。改写后的损失函数有效解决了预测值函数爆炸的问题。实验结果表明，UWAC 在专家数据上的试验效果较好。由此可知，不确定性估计依赖于数据质量，加权学习方法仅可以解决策略估计的准确度问题，并不能有效处理策略估计不好的问题。

不同于以上算法，An 等人[102] 提出了 SAC-N 和集成多样化的行动者评论家(Ensemble-Diversified Actor-Critic，EDAC)算法。SAC-N 在 SAC 算法基础上，采用集成的方式将截断双 Q 学习(Clipped Double Q-Learning)方法中的 Q 网络数量从 2 个提升到了 N 个，并选择 Q 集成中的最小值。也就是说，通过惩罚高不确定性的状态-动作对，从而鼓励策略选择分布内数据的动作。不确定性越高，算法的 Q 值估计方差相应越高。也正是因为分布内数据的样本方差低于 OOD 的样本，可以将这种差异性称为认知不确定性(Epistemic Uncertainty)。为了减

少 SAC-N 算法中所需 Q 网络的数量，EDAC 算法计算集成中每个 Q 函数的梯度，多样化梯度来保证对 OOD 动作有足够的惩罚。同时，采用余弦相似度来衡量评论家的价值网络梯度两两之间的相似程度，并且称其为集成相似度(ES)度量。使用梯度下降法更新评论家：

$$\nabla_{\phi_i}\frac{1}{|\mathcal{D}|}\sum_{(s,a,r,s')\in\mathcal{D}}((Q_{\phi_i}(s,a)-r-\gamma\min_{j=1,2,\cdots,N}Q_{\phi_i}(s',a')+\gamma\beta\log\pi_\theta(a'|s'))^2+\frac{\eta}{N-1}\sum_{1\le i\ne j\le N}ES_{\phi_i,\phi_j}(s,a)) \quad (18)$$

从实验结果来看，与 REM[88] 相比，EDAC 需要的集成网络数量减少了 90%，且在各种类型的数据集上表现出先进的性能。EDAC 既无需对数据收集策略进行显式估计，也不用从分布外数据中进行采样，有效地减少了量化和惩罚认知不确定性所需的集成网络数量，是当前先进的离线 RL 方法之一。

类似地，Bai 等人[103] 提出一种悲观自举的离线强化学习(Pessimistic Bootstrapping for Offline RL，PBRL)算法。该算法采用集成的方式，使用自举 Q 函数，将不确定性度量表示为 Q 函数的标准差：

$$\mathrm{Std}[Q^k(s,a)]=\sqrt{\frac{1}{K}\sum_{k=1}^{K}(Q^k(s,a)-\bar{Q}(s,a))^2} \quad (19)$$

此外，根据数据来源，将训练过程分为两种方式。对于从离线数据集分布内采样状态-动作对，训练目标中的不确定性是基于目标 Q 网络的。而对于从分布外采样的数据，则提供了一种基于预测 Q 网络的伪目标函数，其中包含相应的不确定性信息。由此重构价值网络的损失函数：

$$L(\phi)=\mathbb{E}_{(s,a,r,s')\sim\mathcal{D}_{\mathrm{in}}}[(\hat{T}^{\mathrm{in}}Q^k-Q^k)^2]+\mathbb{E}_{s^{\mathrm{out}}\sim\mathcal{D}_{\mathrm{in}},a^{\mathrm{out}}\sim\pi}[(\hat{T}^{\mathrm{out}}Q^k-Q^k)^2] \quad (20)$$

策略网络损失函数仍通过最大化 K 个 Q 函数的最小值得到。实验结果表明，PBRL 算法在次优数据上也有较好的表现。同时由于在 PBRL 算法中添加了随机先验函数，所以丰富了集成的多样性，提升了函数的泛化能力。

在 PBRL 算法基础上，Yang 等人[104] 提出了鲁棒离线强化学习(Robust Offline Reinforcement Learning，RORL)算法。该算法对离线数据集附近的状态进行策略和值函数的正则化，并通过最小化 Q 值差来强制每个 Q 函数的平滑性：

$$L_{\mathrm{smooth}}(\phi)=\min_{\hat{s}\in\mathcal{B}(s,\epsilon)}-L(Q_\phi(\hat{s},a),Q_\phi(s,a))=\min_{\hat{s}\in\mathcal{B}(s,\epsilon)}-((1-\tau)\delta(s,\hat{s},a)_+^2+\tau\delta(s,\hat{s},a)_-^2) \quad (21)$$

其中，$\delta(s,\hat{s},a)=Q_\phi(\hat{s},a)-Q_\phi(s,a)$，如果 $\delta(s,\hat{s},a)$ 大于 0，则需要平滑过高估计的 Q 值函数，相反 $\delta(s,\hat{s},a)$ 小于 0，则 Q 值函数将被低估。τ 为控制平衡的超参数，$\mathcal{B}(s,\epsilon)$ 是一个对抗性的状态数据集。此外，该算法对这些 OOD 状态的保守值进行了额外估计。实验结果表明，RORL 算法不仅有效平衡了离线 RL 的鲁棒性和保守性，而且对不同类型攻击带来的对抗性扰动具有较强的鲁棒性。

基于策略梯度的 AC 算法是一种最基础的 DRL 算法，它有效地结合了策略网络与价值网络。同时，基于策略梯度的 AC 算法对于离线 RL 来说也是非常重要的。不仅可以帮助智能体更好地利用离线数据，还能提高学习的效率和稳定性。因此，策略梯度法往往是最简单、最通用和最有效的离线 RL 方法。

4.4 对比分析

无模型离线 RL 方法具有易于实现和理解的优势。根据方法的原理不同，无模型离线 RL 方法分为基于表征学习、模仿学习和策略梯度的方法。近几年的研究表明，在无模型离线 RL 算法中，基于表征学习的算法占 48.4%，基于策略梯度的算法占 35.5%，而基于模仿学习的算法占 16.1%。这表明离线 RL 与表征学习相结合的研究方向受到显著的关注，传统的策略梯度方法仍占据着重要地位，而离线 RL 与模仿学习相结合的方法有待进一步扩展。

总体来看，无模型离线 RL 的研究已经有了初步成效，但是仍需要进一步研究和探索。表 1 对本节涉及的方法以表格形式进行描述，包括其主要创新点和缺陷。表 2 对本节提及算法的年份和实验环境以表格形式进行补充说明，符号“•”表示算法使用了某一环境，具体涉及的 D4RL 和 RL Unplugged 环境将在第 7 节展开描述。其中，表 2 涉及的其他实验环境包括：BMA 算法所使用的现实应用中的推荐系统和对话系统任务；KFC 算法所使用的多任务环境[75] (MetaWorld)和机器人模拟框架[76] (RoboSuite)；GSF 算法所使用的离线版本的 Procgen[78] 环境；ABM-MPO 算法所使用的一系列 Sawyer 机械臂操纵放置任务；POR 算法所使用的多房间格子世界环境[105]；AWR 算法所使用的运动模拟框架[106]；AWAC 算法所使用的机械臂打开抽屉任务(Sawyer Drawer)、机械臂手部操控任务(Sawyer Hand)和 ROBEL[107] 环境中的旋转物体至目标任务(D'Claw Turn)。

表 1　无模型离线 RL 算法分析与对比

算法类型		算法名称	创新点	缺　陷
无模型	表征学习	BCQ[11]	离线 RL 领域的开创性工作，提出了外推误差，还通过 VAE 和扰动模型来约束策略	对于 OOD 的动作无法很好地拟合，并且只与在线 DQN 算法性能相当
		离散 BCQ[52]	在 BCQ 基础上，提出了一种适用于离散动作空间的算法，并与经典的 RL 算法进行对比，以验证其优越性	只针对 ALE 离散实验环境，可看成是 BCQ 算法的进一步完善
		BEAR[60]	使用基于采样的 MMD 散度，并引入支持匹配的概念，将策略动作限制在训练分布的支持集范围内	算法过于保守，且在 D4RL 实验环境上，其算法性能不佳
		SPOT[61]	引入了一个基于密度的正则化项，可以与其他异策略 RL 方法相结合	在 Gym-MuJoCo 和 AntMaze 任务下的性能提升并不明显
		PLAS[63]	隐式地将策略限制在数据集的支持范围内，并使用 CVAE 对行为策略进行重建	过于依赖数据集的质量与多样性
		LAPO[64]	提出一种潜在空间优势加权策略优化算法，为多模态离线 RL 提供了新的思路	使用预先训练的 VAE，应确保它的训练数据能充分覆盖高回报样本的分布
		BMA[65]	提出了一种基于伪度量的离线 RL 动作表示学习框架	实验参数过多，整体框架较为复杂
		MCQ[66]	在不损害值函数泛化性的情况下，提出了轻度保守 Q 学习算法	实验不充分，只在 Gym-MuJoCo 任务下进行了验证
		ORL-MI[67]	提出一种基于互信息的离线 RL 动作嵌入模型，理论上可以提高值函数的泛化能力	实验部分未能清晰展示动作是如何进行泛化的
		CNF[69]	利用保守标准流模型构建了一个潜在的动作空间，允许潜在策略使用整个潜在空间	无明显缺陷
		Diffusion-QL[70]	引入条件扩散生成模型，可以更好地捕捉多模态分布	超参数数量增加，需要更多的计算资源进行训练
		SGBCQ[73]	使用 GAN 代替 BCQ 算法中的 VAE，并将安全限制条件融入到离线 RL 中	GAN 容易引起模式塌缩问题，且与后续算法对比性能一般
		KFC[74]	提出了一种基于对称性的数据增强技术，该技术源自库普曼潜在空间表示	无法应对不连续的任务
		GSF[77]	将 RL 与表示学习相结合，以提高零样本在基于像素的控制任务中的泛化性	实验结果过于依赖超参数的选择
		BPR[79]	提出了一种状态表示学习的算法，该算法可以与现有的离线 RL 算法相结合	当训练环境与评估环境存在显著差异时，其泛化性需要进一步提高
	模仿学习	BAIL[13]	将状态值函数拟合到静态数据集的近似上包络，近似出满足贝尔曼方程的最优值函数	要求数据集的覆盖性较好，在迷宫世界或者 ALE 环境中产生限制
		ABM-MPO[14]	将 ABM 作为 MPO 的先验方法并滤除导致性能不如当前策略的轨迹	没有在相同的情况下进一步验证该算法是否适用于多模态环境
		CRR[15]	使用优势函数的估计值在静态数据集中为 BC 选择最佳动作，并扩展到高维状态-动作空间	没有在 Gym-MuJoCo 任务下进行实验
		RvS[16]	采用条件策略模型来学习策略，并且提出了一种通过监督学习实现的方法	算法在随机数据上的性能不佳，且模型性能对于条件变量取值较为敏感
		POR[17]	通过解耦的学习方式，实现了离线数据集的状态拼接	在 Gym-MuJoCo 和 AntMaze 任务下的性能提升并不明显
	策略梯度	BRAC[89]	基本上为策略约束方法构造了一个比较完备的算法框架，并对超参数问题进行讨论	创新性不足，且在 D4RL 实验环境中，该算法性能不佳
		TD3＋BC[91]	将 TD3 与 BC 相结合，旨在解决如何简单并高效地实现算法	算法理论性不足，更多的是简化过程与提高样本学习效率
		PRDC[92]	对数据集进行约束，解决了 Q 值函数过高估计的问题	对超参数的依赖性较强
		AWR[94]	对 RWR 算法进行改进，将策略优化看成是极大似然估计问题，提出一种隐式约束方法	在 RWR 算法上的改进缺少创新性
		AWAC[95]	实现离线预训练在线微调，从而提高强化学习算法本身的训练效率	该算法性能落后于部分离线 RL 方法
		One-step[96]	创新性地提出单步约束思想，解决了离线 RL 多步中遇到的迭代误差问题	在大部分任务下，单步约束算法性能远不如多步的有效
		CQL[12]	在 Q 值函数上添加正则化项，并从理论上证明了可以产生一个当前策略真实值的下界	在 Adroit 动作空间庞大的任务下，该算法性能一般
		Fisher-BRC[98]	用适当的正则化偏移项来增加标准的贝尔曼误差评论家函数损失，同时引入 Fisher 信息距离	对正则化系数的适合范围要求比较严格
		IQL[100]	无需对 OOD 动作进行任何查询，还可以执行多步动态规划	该算法性能落后于部分离线 RL 方法
		UWAC[101]	在 BEAR 算法的基础上引入对状态-动作对的不确定性估计，并将估计的不确定性融入行动者和评论家的损失函数中	过于依赖数据质量，与之后离线 RL 算法相比，该算法性能不佳
		EDAC[102]	采用余弦相似度来衡量价值网络梯度两两之间的相似程度，并减少了量化认知不确定性所需的集成网络数量	该算法需要的集成网络数量仍大于 PBRL 算法
		PBRL[103]	采用集成方式，提出 OOD 采样方法，并用 Q 函数估计的标准差表示不确定性估计	该算法性能不如 EDAC 算法
		RORL[104]	采用保守平滑技术，不仅平衡了离线 RL 的保守性和鲁棒性，而且对不同类型攻击带来的对抗性扰动具有较强的鲁棒性	在采样对抗性状态时，该算法的速度较慢

表 2 无模型离线 RL 算法的实验环境

算法名称	年份	D4RL				RL Unplugged			其他
		Gym-MuJoCo	Adroit	Franka Kitchen	迷宫世界	DM 运动	DM 控制	ALE	
BCQ[11]	2019	●							
离散 BCQ[52]	2019							●	
BEAR[60]	2019	●							
SPOT[61]	2022	●			●				
PLAS[63]	2021	●	●	●	●				
LAPO[64]	2022	●		●	●				
BMA[65]	2022				●				●
MCQ[66]	2022	●							
ORL-MI[67]	2022	●							
CNF[69]	2022	●			●				
Diffusion-QL[70]	2022	●	●	●	●				
SGBCQ[73]	2023	●							
KFC[74]	2022	●	●	●	●				●
GSP[77]	2022								●
BPR[79]	2023	●					●		
BAIL[13]	2020	●							
ABM-MPO[14]	2020	●							●
CRR[15]	2020					●	●		
RvS[16]	2022	●		●	●				
POR[17]	2022	●			●				●
BRAC[89]	2019	●							
TD3+BC[91]	2021	●							
PRDC[92]	2023	●			●				
AWR[94]	2019	●				●		●	●
AWAC[95]	2020	●	●						●
CQL[12]	2020	●	●	●	●			●	
Fisher-BRC[98]	2021	●							
One-step[96]	2021	●	●						
IQL[100]	2021	●	●	●	●				
UWAC[101]	2021	●	●						
EDAC[102]	2021	●	●						
PBRL[103]	2022	●	●						
RORL[104]	2022	●							

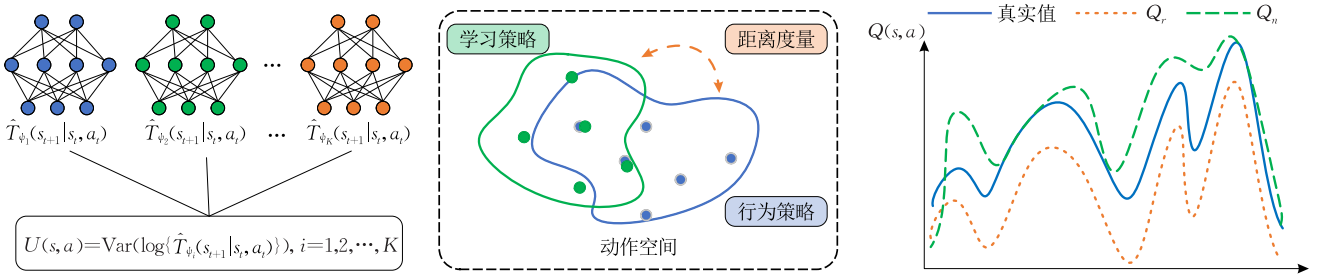
5 基于模型

在 RL 领域中，环境 MDP 模型主要包含状态、动作、状态转移模型与奖励函数。基于模型的 RL 可以通过经验数据直接模拟真实环境，并且能与监督学习相结合来求解环境模型，因此被认为是解决现实世界序列决策问题的一个有效方向[108]。但由于环境模型中的策略推演需要不断在模型预测的基础上进行进一步的预测，而这通常伴随着复合误差，导致最终的算法性能有很大的误差。

与基于模型的在线 RL 方法相似，基于模型的离线 RL 方法首先需要从数据集中学习状态转移模型 $\hat{T}(s'|s,a)$和奖励函数 $r(s,a)$，将其作为真实环境并模拟转移，再通过规划生成动作。但是，一个核心问题是数据集学到的状态转移模型是针对行为策略 π_β 的，并非所学到的策略 π。所以，如何解决基于模型方法中的分布偏移仍是一个公开的挑战性问题。因此，对于离线 RL 而言，设计一种能在给定的分布范围内采取动作、并在其之外获得良好性能的算法是必要的。基于模型的离线 RL 可有效解决此问题，一方面模型本身是离线数据分布的自然扩展，另一方面模型的构造解决了离线 RL 的分布转移问题。但是，数据的分布偏移仍然可能造成环境模型的误差。总之，对于基于模型的离线 RL 方法而言，在离线数据中构建环境模型的关键优势是利用模型的泛化能力来执行一定程度的探索。并且生成额外的训练数据以提高策略性能，这是无模型的离线RL方法远远无法达到的。

根据近几年基于模型的离线 RL 方法的技术发展路线，从不同方法原理出发，将其细分为不确定性估计（Uncertainty Estimation）、策略约束（Policy Constraints）与值函数正则化（Value Function Regularization）共 3 个子类别。对于不确定性估计而言，根据高低不同的不确定性区域，采取不同的惩罚或奖

励。策略约束则是通过对学习策略进行约束或施加惩罚，使其不去或尽可能少地访问 OOD 状态或动作。值函数正则化在值函数上添加正则项，以停止或降低对于 ODD 状态或动作的价值估计。为了直观地对比 3 个类别的区别，绘制图 5 如下。其中，(a)使用模型的集成来估计不确定性；(b)通过度量行为策略与当前学习策略的距离，显式地估计行为策略；(c)比较经过正则化(Regularization)和天真(Naive)的 Q 值。对于具有正则化的方法而言，不易出现过高估计 Q 值的问题。



(a) 不确定性估计　　(b) 策略约束　　(c) 值函数正则化

图 5　基于模型离线 RL 的 3 种方式

5.1　不确定性估计

不确定性估计已成为基于模型的 RL 的常用技术之一。通过该方法，已有基于模型的异策略 RL 算法[109-110]表现出优异的性能，其中部分方法允许在线进行额外的数据收集，但同时也使用了离线数据。基于不确定性的离线 RL 方法允许根据对模型泛化性的信任程度，在保守和天真的离线 RL 方法之间进行切换。通过将不确定性作为正则项的方法，达到限制学习策略的目的。可以在高不确定性区域令策略趋于保守，在低不确定性区域放松对策略的约束。为了避免模型的分布发生偏移，可以估计一个保守的模型来减少 OOD 动作的产生。在这些 OOD 状态下，模型的惩罚奖励函数表示为

$$\tilde{r}(s,a)=\hat{r}(s,a)-\lambda U_r(s,a) \tag{22}$$

其中，λ 为惩罚系数，$U_r(s,a)$是状态-动作对的不确定性度量，并期望数据集中存在的不确定性度量较低。此外，可以将模型中离线 RL 探索置信度问题转化为特定区域的模型不确定性问题[111]。在低维 MDP 中，这种不确定性估计可以通过高斯过程产生[112]，而对于高维问题，可使用贝叶斯神经网络[113]等方法。

作为基于模型离线 RL 的开山之作，将不确定性视为风险与收益的均衡，其中风险指不准确区域过度拟合导致的动态误差，收益指通过在模型的这些区域进行探索而获得的潜在奖励。为了权衡这两个方面，Yu 等人[18]提出了一个基于模型的离线策略优化(Model-Based Offline Policy Optimization，MOPO)算法，通过引入不确定性惩罚奖励(Uncertainty-Penalized Reward)来约束其模型，以使具有保守估计的收益最大化。并给出状态转移模型为 $\hat{T}_{\psi_i}(\cdot|s,a)=\mathcal{N}(\mu_i(s,a),\Sigma_i(s,a))$。则不确定性被定义为 $U_r(s,a)=\max_i\|\Sigma_i(s,a)\|_F$，其中，$\|\cdot\|_F$ 为 Frobenius 范数。由此得到 MOPO 的实际不确定性惩罚奖励为 $\tilde{r}(s,a)=\hat{r}(s,a)-\lambda\max_{i=1,2,\cdots,N}\|\Sigma_i(s,a)\|_F$。其中，$\hat{r}$ 为 $\hat{T}$ 引导的平均预测奖励。有了新的奖励函数后，就可以使用经典的基于模型的方法来解决离线 RL 问题。例如，基于 Dyna 模型[114]的数据扩充强化学习[115-116]、使用 LQR[117]和 PETS[113]等方法在模型下进行轨迹优化或规划以及通过 MVE[118]等方法改进价值函数的评估等。最终，从实验结果来看，在 D4RL 实验环境上 Gym-MuJoCo 任务中，MOPO 算法优于无模型算法。此外，该算法在需要分布外泛化的 HalfCheetah 与 Ant 任务中也表现出了显著的性能提升。并且成功地将基于模型的方法应用到离线 RL 中，不用再依赖于策略约束与正则化等无模型方法。与 MOPO 算法相似，Kidambi 等人[19]提出一种基于模型的离线强化学习(Model-Based Offline Reinforcement Learning，MOReL)算法。该算法从数据集中学习悲观 MDP(Pessimistic-MDP，P-MDP)，并将其用于策略搜索。P-MDP 将状态空间划分为“已知”和“未知”区域，并对未知区域使用大量的负奖励作为惩罚。此外，MOReL 算法通过模型集成来衡量其模型的不确定性，基于此，可以得到状态转移模型为 $\hat{T}_{\psi_i}(\cdot|s,a)=\mathcal{N}(f_{\phi_i}(s,a),\Sigma)$。通过对函数 f_{ϕ_i} 进行参数化来确保局部的连续性 $f_\phi(s,a)=s+\sigma_\Delta\mathrm{MLP}_\phi((s-\mu_s)/\sigma_s,(a-\mu_a)/\sigma_a)$。则不确定性的度量公式为

$$U_r(s,a)=\begin{cases}r_{\max}, & \mathrm{disc}(s,a)>threshold\\ 0, & \text{其他}\end{cases} \tag{23}$$

$$\mathrm{disc}(s,a)=\max_{i,j}\|f_{\phi_i}(s,a)-f_{\phi_j}(s,a)\|_2 \tag{24}$$

其中，disc(s,a)用来衡量状态转移模型之间的不一致。实验表明，在 D4RL 实验环境上 Gym-MuJoCo 任务中，MOReL 算法具有良好的性能，并优于最先进的 CQL[12]算法与 MOPO 算法。但是为了更好的泛化保证，该算法要求数据的可用性较强以及表征质量应该较高。

Rafailov 等人[20]提出了基于潜在离线模型的策略优化(Latent Offline Model-Based Policy Optimization，LOMPO)算法。类似地，该算法依赖于量化模型预测中的不确定性能力，将离线 RL 扩展到高维视觉观察空间，但是这对于视觉观察极具挑战性。使用离线数据学习一个具有图像编码器、解码器和潜在状态转移模型 $\hat{T}_{\psi_i}(\cdot\mid s,a)$集成的变分模型，并且基于潜在状态空间的前向模型之间的分歧来量化不确定性：

$$U_r(s,a)=\mathrm{Var}(\{\log\hat{T}_{\psi_i}(s_t\mid s_{t-1},a_{t-1})\}),\quad i=1,2,\cdots,K \tag{25}$$

则 LOMPO 的实际不确定性惩罚奖励为

$$\tilde{r}(s,a)=\frac{1}{K}\sum_{i=1}^{K}\hat{r}(s^{(i)},a)-\lambda U_r(s,a) \tag{26}$$

其中，$s^{(i)}\sim\hat{T}_{\psi_i}(s_{t-1},a_{t-1})$是从每个前向模型中采样得到的，$s$ 是从 $s(i)$，$i=1,2,\cdots,K$ 中采样得到的。实验结果表明，在 4 个视觉任务与一个现实世界机器人操纵任务中，LOMPO 算法全面优于或匹配先前基于模型的离线 RL 方法。在未来工作中，可以将 LOMPO 算法应用于多任务实验中。通过使用来自多个任务的数据来训练模型，为此学习到更加准确的状态转移模型，从而全方面提高采样效率与泛化能力。

针对当前基于模型的离线 RL 方法中存在的模型回报与其不确定性之间的权衡问题，Yang 等人[21]提出了帕累托策略池(Pareto Policy Pool，P3)算法。该算法将回报和不确定性视为需要平衡的双目标，从而构建一个双目标任务。通过使用帕累托策略池存储策略，P3 能够有效地在未知环境中选择合适的策略。研究中还提出了帕累托推理方法，以降低策略不断扩充的成本代价。实验结果表明，在 D4RL 环境的 Gym-MuJoCo 任务中，P3 算法显著优于 MOPO 和 MOReL 算法。总的来说，P3 是一种策略集成型的多任务双目标优化方法。

5.2 策略约束

策略约束可以根据是否直接估计行为策略分为显式(Explicit)策略约束与隐式(Implicit)策略约束。其中，显式策略约束离线 RL 方法通过约束当前的学习策略 π，使其尽可能地逼近行为策略 π_β。隐式策略约束离线 RL 方法不依赖于对行为策略 π_β的估计，不仅可以使用修正的目标函数对 π 进行约束，还可以基于状态从潜在空间映射到动作空间。

不同于MOPO中采用P-MDP的方法，Matsushima 等人[22]提出行为正则化模型集成(Behavior-Regularized Model-Ensemble，BREMEN)算法。该算法通过适当的策略初始化和信任区域更新显式地约束策略使其接近行为策略。BREMEN 使用 K 个状态转移模型 $\hat{T}(\cdot\mid s,a)$的集合，以缓解模型偏差问题。通过 ϕ_i参数化每一个模型$\hat{f}_{\phi_i}$，来对目标函数进行训练。这使得在数据集$\mathcal{D}$上，下一个真实状态 s'与下一个预测状态$\hat{f}_{\phi_i}(s,a)$之间的均方误差最小化。为了使学到的策略接近数据收集的行为策略，选择使用基于 KL 散度的信任区域优化。因此，最大化目标函数来优化策略：

$$J(\theta_{k+1})=\arg\max_{\theta}\mathbb{E}_{s,a\sim\pi_{\theta_k},\hat{f}_{\phi_i}}\left[\frac{\pi_\theta(a\mid s)}{\pi_{\theta_k}(a\mid s)}A^{\pi_{\theta_k}}(s,a)\right]$$
$$\text{s.t.}\ \mathbb{E}_{s\sim\pi_{\theta_k},\hat{f}_{\phi_i}}[D_{\mathrm{KL}}(\pi_\theta(\cdot\mid s)\|\pi_{\theta_k}(\cdot\mid s))]\leq\delta \tag{27}$$

其中，初始化目标策略 π_θ为均值为 π_β、标准差为 1 的高斯策略。δ 为最大步长，$A^{\pi_{\theta_k}}(s,a)$为优势函数。BREMEN 算法的优点是提出了一个新的衡量离线 RL算法的标准，即部署效率(Deployment Efficiency)。它可以计算出在学习期间数据收集策略的变化次数。实验结果表明，BREMEN 可以使用比现有算法少 10～20 倍的数据，有效地优化策略。并在有限的部署环境中显示出良好的结果，例如，在 Gym-MuJoCo 任务中，从开始学习到成功的策略只需 5～10 次部署，而对于传统 RL 算法来说，可能需要付出几百倍甚至几百万倍的代价。这不仅可以减轻现实世界应用中的成本和风险，还可以减少分布式学习过程中所需要的通信量。但是，该算法的稳定性不佳，有时不如没有部署约束的 SAC 算法。

5.3 值函数正则化

不同于策略约束的方法，值函数正则化不对行为策略进行限制，故该方法通常没有策略约束那样保守。在部分论文中也将值函数正则化称为策略正则化或正则化，其主要思想都是在目标函数上添加一个正则化项。在无模型离线RL中，CQL 算法可被看作是一种值函数正则化的方法。基于此，Yu等人[23]提出基于保守离线模型的策略优化(Conservative Offline Model-Based Policy Optimization，COMBO)。该算法学习一个状态转移模型 $\hat{T}(\cdot\mid s,a)$，作为下一个状态的高斯分布，并利用最大对数似然估计进行奖励函数的训练。基于此，COMBO引入了一个新的 MDP。在策略评估阶段，最小化损失函数：

$$L(\phi)=\arg\min_{Q}\beta(\mathbb{E}_{s,a\sim\rho(s,a)}[Q_\phi(s,a)]-\mathbb{E}_{s,a\sim\mathcal{D}}[Q_\phi(s,a)])+\frac{1}{2}\mathbb{E}_{s,a,s'\sim d_f^\pi}[(Q_\phi(s,a)-(\hat{\mathcal{B}}^\pi\hat{Q}_\phi^k)(s,a))^2] \tag{28}$$

其中，$\hat{\mathcal{B}}^\pi\hat{Q}_\phi^k(s,a)=r(s,a)+Q^\pi(s',a')$，$s'\sim\mathcal{D}$，$a'\sim\pi(\cdot|s')$是基于样本的贝尔曼算子。当模型产生了优异的状态-动作对，并且与真实状态-动作对无法区分时，式(28)中的两个正则化项可以进行平衡。Trabucco 等人[119]指出可将这种正则化视为类似对抗性训练。通过惩罚 OOD 的状态-动作对，希望生成器能够达到优秀水平。此外，$\rho(s,a)$和 d_f^π是可以选择的采样分布，$f\in[0,1]$是静态数据集中的数据点比率，可以将其视为保守程度。f 值越大，表示在静态数据集中进行的采样越多。因此，将获得更加保守的 Q 值估计，反之亦然。从实验结果来看，在 3 个需要适应不可视行为的任务中，COMBO 算法展现出良好的泛化性和稳定性。此外，该算法在不同类型的数据集上均表现出色，这表明其对不同类型的数据集具有鲁棒性。由此可将 COMBO 算法视为正则化与基于模型方法的一种结合形式。与无模型的离线 RL 方法相比，COMBO 算法中的 Q 值函数估计不会过于保守，并且能够确保策略的稳定提升。对比以上基于模型的离线 RL 方法，COMBO 算法有效消除了不确定性对神经网络拟合带来的影响。但是，该算法无法在不同的数据集中统一超参数，且没有考虑如何根据模型误差自动选择 f。

5.4　对比分析

基于模型的方法可以充分利用样本逼近模型，使得数据利用率极大提高[120-121]。因此，对于离线 RL 而言，该方法可以有效利用离线数据，从而展现出无模型方法无法超越的优势。同时，模型通常对环境的变化具有鲁棒性，即使遇到新环境，算法也能够依靠已学到的模型进行推理，具有很好的泛化能力。但是，基于模型的方法需要学习状态转移模型，在实际情况中，这是一项极具挑战性的工作。此外，该方法在高维空间与长时段问题上仍存在困难。

近几年的研究表明，在基于模型的离线 RL 算法中，基于不确定性估计的算法占 66.6%，基于策略约束和值函数正则化的算法均占 16.7%。这表明对环境不确定性的考虑在算法设计中扮演着重要角色。尤其是基于不确定性估计的方法逐渐成为研究的主流方向，且在该分类下，LOMPO 算法明显优于其他算法。值得注意的是，这些方法在解决基于模型离线 RL 问题时，通常不是相互孤立的，而是经常交叉使用以增强方法的效果。

表 3 对本节涉及的算法以表格形式进行描述，包括其主要创新点和缺陷。表 4 对本节提及算法的年份和实验环境以表格形式进行补充说明，符号“●”表示算法使用了某一环境。具体涉及的 D4RL 和 RL Unplugged 环境将在第 7 节展开描述。其中，表 4 涉及的其他实验环境包括：LOMPO 算法所使用的机械臂开门任务(Sawyer Door Open)、真实机

表 3　基于模型的离线 RL 算法分析与对比

算法类型		算法名称	创新点	缺陷
基于模型	不确定性估计	MOPO[18]	将基于模型的方法应用到离线 RL 中，并引入不确定性惩罚奖励来限制策略的学习	与之后基于模型的离线 RL 算法相比，其算法性能不佳
		MOReL[19]	通过模型集成来衡量模型的不确定性，并使用数据集学习 P-MDP	对数据和表征的质量要求较高
		LOMPO[20]	引入变分模型，将潜在模型不确定性作为潜在模型集成不一致来惩罚潜在状态	无明显缺陷
		P3[21]	开发了一种有效的方法，该方法在帕累托前沿执行不同层次的权衡，为推理阶段选择最佳策略提供了灵活性	在实验部分，所选择的实验环境与对比算法均不够充分
	策略约束	BREMEN[22]	提出部署效率作为一个新的衡量 RL 算法的标准，可以使用比现有算法少 10～20 倍的数据，有效地优化策略	算法稳定性不佳，有时不如没有部署约束的 SAC 算法
	值函数正则化	COMBO[23]	提出了一种避免不确定性估计的方法，将 CQL 算法与基于模型相结合，使 Q 值函数估计不宜过于保守	没有考虑到如何根据模型误差自动选择比率 f

表 4　基于模型离线 RL 算法的实验环境

算法名称	年份	D4RL				RL Unplugged			其他
		Gym-MuJoCo	Adroit	Franka Kitchen	迷宫世界	DM 运动	DM 控制	ALE	
MOPO[18]	2020	●							
MOReL[19]	2020	●							
LOMPO[20]	2021	●	●				●		●
P3[21]	2022	●							
BREMEN[22]	2021	●							
COMBO[23]	2021	●							●

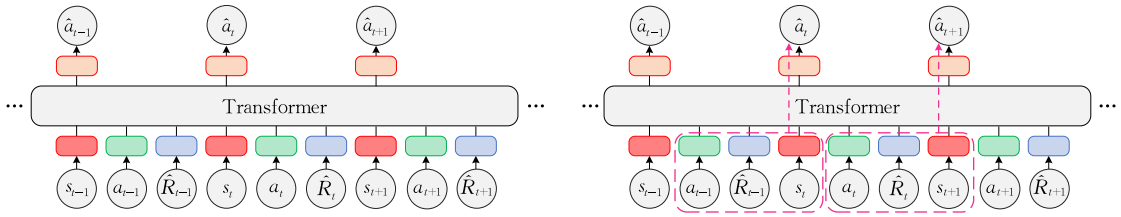
器人环境和 ROBEL[107] 环境中的连续旋转物体任务(D'Claw Screw);COMBO 算法所使用的机械臂关门任务(Sawyer Door Close)。

6 基于 Transformer 模型

Transformer[122] 具有强大的表征能力和时序建模能力,其最早提出用于自然语言处理领域,可以对语义概念的高维分布进行规模化建模,包括语言中有效的零点泛化[123] 和图像生成[124]。自然语言处理领域的大多数问题是序列问题,RL 本质上也是一个序列决策问题。因此,创新性地将 Transformer 运用在 RL 中。该方法不同于传统无模型 RL 采用 MDP 建模,其运用基于值函数方法如时序差分(TD)或参数化策略方法如策略梯度(PG)进行动作选择。将 RL 问题看成一个序列生成任务,通过神经网络直接输出动作。

在基于 Transformer 模型的离线 RL 中,关注的是如何在整个轨迹上训练一个联合状态-动作模型,并选择出最佳动作。因为在整个轨迹中存在多个状态和动作的锚点(Anchor),来防止学到的策略偏离 π_β 太远,所以序列建模使得智能体不容易产生 OOD 动作,无需通过约束或悲观假设来解决外推误差的问题。此外,因为离线 RL 无需在线收集数据和更新模型的特点,使得 Transformer 模型在处理长序列问题上有很好的效果。因此,基于 Transformer 模型的离线 RL 方法,彻底解决了分布偏移的问题,从而将目标聚焦在解决 RL 中的长时序(延迟奖励)与稀疏奖励等经典问题上。

根据近几年基于 Transformer 模型的离线 RL 技术发展路线,从轨迹序列出发,将基于 Transformer 模型的框架分为回报导向型(Return-To-Go)与元组导向型(State-Action-Return)。在回报导向型中,将轨迹序列表示为长度为 T 的状态-动作序列 $\tau=s_1,a_1,\hat{R}_1,s_2,a_2,\hat{R}_2,\cdots,s_T,a_T,\hat{R}_T$。其中,$\hat{R}_t$ 与传统 MDP 建模不同,这里指状态-动作序列在 t 时刻后的所有奖励之和即回报,而非传统 MDP 中的即时奖励。然而,在元组导向型中,考虑了近期的未来状态和相应的奖励是当前动作的直接结果。因此,在回报导向型的基础上,将一条轨迹分解为多个状态-动作-回报元组。图 6 对基于 Transformer 模型离线 RL 的两种方法进行了直观地对比。



(a)回报导向型 (b)元组导向型

图 6 基于 Transformer 模型离线 RL 的两种方式

6.1 回报导向型

回报导向型是一种通过建模状态、动作与回报之间的关系来实现离线 RL 的方法。以回报为先验知识,输入当前状态后算法根据目标函数找出最有可能的动作,这是先前两类离线 RL 方法无法做到的。

Chen 等人[24] 提出一种基于 Transformer 模型的决策转换器(Decision Transformer,DT)架构,可将离线 RL 问题转换为条件序列建模。该方法基于回报导向型,使用 GPT 架构模拟轨迹分布。在训练阶段,最小化真实动作与预测动作之间的均方误差。在评估阶段,根据任务的期望性能,对轨迹进行调节并更新其回报以达到目标。实验表明,在 ALE、Gym-MuJoCo 和 Key-To-Door[125] 这 3 个任务中均取得了较好的成绩。虽然训练成本很高,但在解决稀疏奖励与延迟奖励的问题上,DT 的表现远超过了最先进的无模型离线 RL 算法,例如 CQL[12]。然而,对于其他回报导向型方法而言,DT 的泛化能力不强,且只适用于离线 RL,无法将其扩展到在线学习中。

类似地,Janner 等人[25] 沿用 DT 思想提出一种轨迹转换器(Trajectory Transformers,TT)架构,将 RL 问题看成序列生成任务。与 DT 算法相比,二者都使用了具有自回归生成能力的 GPT 结构来生成轨迹。特别地,TT 结合波束搜索(Beam Search)规划对候选轨迹进行搜索和优化。从实验来看,基于 Transformer 模型的轨迹建模能力远超于基于模型的方法。在 Gym-MuJoCo 任务下的离线 RL 任务与多房间格子世界环境[105] 下的目标条件强化学习[126](Goal-Conditioned Reinforcement Learning,GCRL)任务中,TT 都取得了不错的效果。并且因为其具有可扩展的表征架构,可以有效地对复杂问题进行表

征提取，从而处理 RL 本身无法解决的问题。但与传统的动态规划算法相比，由于 TT 使用最大似然目标训练 Transformer 模型，导致该方法更依赖于训练数据的分布。为了解决 DT 只适用于离线 RL 的问题，Zheng 等人[26]提出在线决策转换器（Online Decision Transformer，ODT）架构，将离线预训练与在线微调进行融合统一。ODT 方法结合了 DT 算法与 AWAC[95]思想，在在线阶段利用了积极的采样与探索，并将其直接适用于 POMDP 过程。实验表明，在 Gym-MuJoCo 任务中，ODT 可大大超越 DT 基线。但在 AntMaze 任务中，ODT 略逊于经过在线微调后的 IQL[100]算法，说明该方法只在部分环境下有良好的效果。

为进一步定义广义 DT 的形式，增强其泛化能力，Furuta 等人[27]提出一种用于解决事后信息匹配（Hindsight Information Matching，HIM）问题的通用决策转换器（Generalized Decision Transformer，GDT）架构。通过引入 HIM 作为现有事后启发（Hindsight-Inspired）算法的统一形式，并将 GDT 作为更为通用的 DT 方法，用序列建模的方式解决任何 HIM 问题。并对回报导向型方法进行推广，将信息匹配（Information Matching）问题定义为一个条件策略 $\pi(a|s,z)$：

$$\min_{\pi}\mathbb{E}_{z\sim p(z),\tau\sim\rho_z^{\pi}(\tau)}[D(I^{\Phi}(\tau),z)] \tag{29}$$

其中，给定状态 s_t 的部分轨迹 $\tau_t=\{s_t,a_t,s_{t+1},a_{t+1},\cdots\}$，将其信息统计（Information Statistics）定义为 $I(\tau)$，则 $I^{\Phi}(\tau)$ 为特征函数 Φ 的信息统计，或者说是特征函数对应轨迹 $\tau_t^{\Phi}=\{\phi_t,\phi_{t+1},\cdots,\phi_T\}$，$\phi_t=\Phi(s_t,a_t)$ 的聚合。对于任何给定的轨迹，让 $z^*=I^{\Phi}(\tau)$ 最小化散度 $D=0$，故当 $z=z^*$ 时，状态-动作序列是最佳的，(τ_i,z_i) 可用于 RL 或 BC，这些算法称为 HIM 算法。当特征函数是奖励，且聚合器为折扣求和，则为 DT 方法；如果聚合器是分类，则得到离线多任务状态边际匹配的分类 DT（Categorical Decision Transformer）方法；如果聚合器是 Transformer，则获得用于离线多任务模仿学习的双向 DT（Bi-Directional Decision Transformer）方法。特征函数 Φ 和聚合器的选择一起决定了 HIM 中的 $I^{\Phi}(\tau)$，也就是说，GDT 可以通过正确选择特征函数 Φ 和聚合器来解决 HIM 问题。

6.2　元组导向型

元组导向型在回报导向型的基础上进行改进，明确地对状态-动作-回报整个元组进行学习。因此，将其看作是一种局部特征建模的方法，这样可以更好地帮助算法完成长期序列建模。

不同于回报导向型，Shang 等人[28]创新地提出一种用于视觉的状态-动作-回报转换器，简称 StARformer 架构。该算法由单步转换器与序列转换器两部分组成，分别对单个步骤与整个序列进行建模。因此，元组导向型与回报导向型的区别在于是否通过单步转换过程来学习每个元组内的局部关系。对于每个单步转换器层 l 的输出端，通过聚合输出标记（Token）Z_t^l 得到状态-动作-奖励表征 g_t^l：

$$g_t^l=\mathrm{FC}([Z_t^l])+e^{\text{temporal}} \tag{30}$$

其中，$[\cdot]$ 表示每个元组内的标记和 e^{temporal} 的串联。为了创建纯状态标记 h_t^0，将状态作为整体嵌入：

$$h_t^0=Conv(s_t)+e^{\text{temporal}} \tag{31}$$

然后，将输出的表征 g_t^l 和 h_t^0 输入到相应的序列转换器层进行长期序列建模。以此得到一个合并的序列：

$$Y_{\text{in}}^l=\{g_1^l,h_1^{l-1},g_2^l,h_2^{l-1},\cdots,g_T^l,h_T^{l-1}\} \tag{32}$$

因为对状态-动作-奖励序列采取不同粒度的建模，并将类似 MDP 的感应式偏差（MDP-Like Inductive Bias）引入模型中，使得模型在相比较下更容易扩展到长序列问题中。从实验结果来看，StARformer 算法在 ALE 环境和 DM 控制套件测试中，都优于其他基于 Transformer 模型的方法。StARformer 可以在不降低性能的情况下实现逐步奖励。然而，在 DT 中需要仔细设计目标返回值，并进行反复的试验和调整才能找到最佳值。

6.3　对比分析

总体来看，基于 Transformer 模型的离线 RL 具有较强的稳定性和一定的泛化能力，专注于解决传统 RL 难以应对的长序列问题。但同时，也不可避免地带来了对实验配置要求较高、需要消耗大量的时间以及资源的问题。对比回报导向型与元组导向型，后者有助于算法更好地处理长期序列建模问题。通过考虑整个元组，准确地捕捉状态和动作之间的关联，从而高效地理解和适应环境。

表 5 对本节涉及的算法以表格形式进行描述，包括其主要创新点和缺陷。表 6 对本节提及算法的年份和环境以表格形式进行补充说明，符号“●”表示算法使用了某一环境。具体涉及的 D4RL 和 RL Unplugged 将在第 7 节展开描述。其中，表 6 涉及的其他实验环境包括：DT 算法所使用的拿到钥匙开门任务[125]（Key-to-Door）；TT 算法所使用的多房间格子世界环境[105]。

表 5 基于 Transformer 模型离线 RL 算法的分析与对比

算法类型		算法名称	创新点	缺陷
基于 Transformer 模型	回报导向型	DT[24]	基于回报,将离线 RL 问题建模成适合于 Transformer 处理的序列问题	没有太多泛化的能力,只适用于离线 RL,无法将其扩展到在线 RL 中
		TT[25]	通过 GPT 架构学习轨迹分布,并使用波束搜索与奖励之和进行规划	与传统离线 RL 对比,对训练数据有较强的依赖性,且预测速度慢
		ODT[26]	将 DT 扩展到在线 RL 中,并且可以直接适用于 POMDP 过程	受限于环境,只在部分环境下有较好的成绩
		GDT[27]	提出了 HIM 和 GDT 框架,为事后算法提供了统一的形式	需要消耗的资源较多
	元组导向型	StARformer[28]	提出状态-动作-回报转换器架构,使得模型更容易扩展到长序列问题	未对不需要输入图像的任务进行实验验证

表 6 基于 Transformer 模型离线 RL 算法的实验环境

算法名称	年份	D4RL				RL Unplugged			其他
		Gym-MuJoCo	Adroit	FrankaKitchen	迷宫世界	DM 运动	DM 控制	ALE	
DT[24]	2021	●						●	●
TT[25]	2021	●			●				●
ODT[26]	2022	●			●				
GDT[27]	2022	●							
StARformer[28]	2022						●	●	

7 实验环境

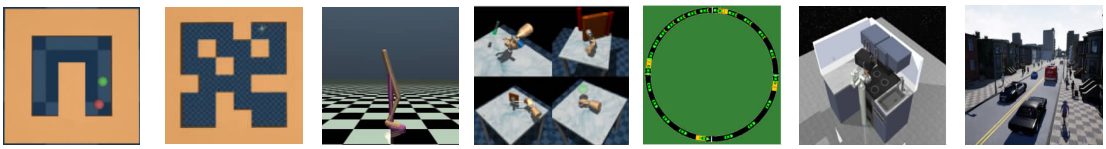
深度学习的成功离不开大规模数据集的支持,这些数据集提供了足够的样本和标签,有助于深度学习算法学习到更加准确、泛化能力更强的模型。同样地,数据集对于离线 RL 也非常关键,它直接决定了离线 RL 算法的性能和效果。本节介绍了迄今为止 3 个最大的离线 RL 实验环境:D4RL[127]、RL Unplugged[128]和 NeoRL[129],概述了每个环境中的任务和属性,并对 3 个实验环境进行对比。

7.1 D4RL

D4RL 实验环境包括 OpenAI 的迷宫世界、Gym-MuJoCo[130]、灵巧操纵任务[131](Adroit)、交通模拟任务[132](Flow)、机器人操纵任务[133](Franka-Kitchen)和自动驾驶任务[134](CARLA)的数据集。

迷宫世界包括 Maze2D 和 AntMaze,可以将它们看成导航任务。这两项任务的不同之处在于在 AntMaze 中,一个更复杂的 8 自由度(8DoF)"蚂蚁"4 边形机器人取代了 Maze2D 中的 2 个小球。Gym[135]作为 OpenAI 的仿真平台,是 RL 中必不可少的开源工具包。其中,Gym-MuJoCo 是一个免费的开源物理引擎,不仅用于实现基于模型的计算,还可作为传统模拟器、模拟游戏和交互式虚拟等环境的实现工具。灵巧操纵任务为控制一个 24 自由度(24DoF)的机械手臂来模拟各项工作。Flow 提供了一个模拟现实世界交通动态的任务,通过控制自动驾驶车辆,最大限度地提高通过环形或合并道路配置的交通流量。机器人操纵任务为控制一个 9 自由度(9DoF)的 Franka 机器人在厨房环境中互动。自动驾驶任务为一个高度真实的自动驾驶模拟器,通过控制汽车的油门、转向和制动踏板来驾驶车辆。

D4RL 实验环境所涉及的部分任务如图 7 所示。(a)为迷宫世界中的 Maze2D;(b)为迷宫世界中的 AntMaze;(c)为 Gym-MuJoCo 中的 Walker2D-v2 任务,代表训练一个双足行走的智能体;(d)为 Adroit 中旋转笔、开门、锤钉子与拾起并移动球体这 4 项操作任务;(e)为 Flow 中控制车辆提高通过环形道路配置的交通流量;(f)为在厨房环境中,控制机器人将微波炉与滑动柜门打开,水壶放在燃烧器上,并把顶灯打开;(g)为 CARLA 中在一个小城镇内的模拟导航。



(a) Maze2D (b) AntMaze (c) Walker2D-v2 (d) Adroit (e) Flow (f) FrankaKitchen (g) CARLA

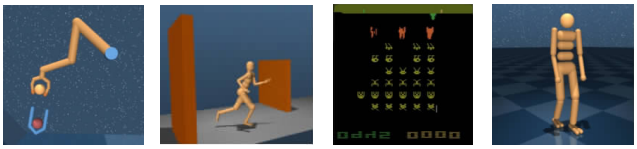
图 7 D4RL 实验环境

7.2　RL Unplugged

RL Unplugged 实验环境由 4 个不同套件的数据集组成，包括 DM 控制套件[136]（DeepMind Control Suite）、DM 运动套件[137]（DeepMind Locomotion Suite）、电玩游戏[138]（Arcade Learning Environment，ALE）和现实世界强化学习套件[139]（Real World RL Suite，RWRL）。

DM 控制套件是一套在 Gym-MuJoCo 中实现的控制任务，包括操纵与位置移动任务。DM 运动套件包括人形走廊移动任务和模拟啮齿动物任务，其特点是将具有高自由度的连续控制与以自我为中心的感知相结合。电玩游戏是一个由 57 个 Atari 2600 游戏组成的多样化套件，旨在通过大型游戏数据集来评估离散的 RL 算法。其作为通用的像素类游戏实验平台，为 DRL 的无模型学习[140]、基于模型规划[141]和模仿学习[142]等研究提供了验证环境[143]。对于现实世界强化学习套件，共评估了 9 个挑战，这些挑战包括高维状态与动作空间的输入和输出、系统延迟、系统约束、多目标、处理非平稳性和部分可观察性。

RL Unplugged 实验环境所涉及的部分任务如图 8 所示。(a)为 DM 控制套件中通过控制操纵器将球放入篮筐任务；(b)为 DM 运动套件中人形穿越走廊（Humanoid Corridor）任务；(c)为 Atari 2600 游戏中太空入侵者（Space Invaders）任务；(d)为现实世界强化学习套件中的人形行走（Humanoid Walk）任务。



(a) Insert ball　(b) Humanoid corridor　(c) Space invaders　(d) Humanoid walk

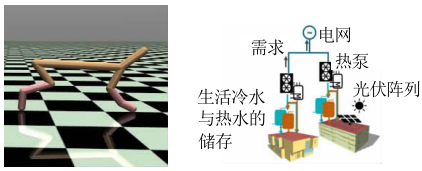
图 8　RL Unplugged 实验环境

7.3　NeoRL

NeoRL 实验环境包括 Gym-MuJoCo 任务、工业基准[144]（IB）、股票交易模拟器[145]（FinRL）和城市管理[146]（CityLearn）的数据集。

该实验环境中的 Gym-MuJoCo 任务与 D4RL 实验环境中类似。工业基准通过模拟各种工业控制任务，来解决现实世界中的相关问题，包括高维连续状态和动作空间、延迟奖励、复杂的噪声模式以及多个反应目标的高随机性。股票交易模拟器可以模仿真实的股票市场，并且环境随着时间推移存在自我演化现象。城市管理是一个类似 Gym 的环境，它通过控制不同类型建筑的储能来重塑电力需求的聚集曲线。

NeoRL 环境所涉及的部分任务如图 9 所示。(a)为 Gym-MuJoCo 中的 HalfCheetah-v2 任务，代表训练一个两足的智能体实现行走；(b)为 CityLearn 中协调建筑物对生活热水与冷水储存的控制流程。



(a) HalfCheetah-v2　(b) CityLearn

图 9　NeoRL 实验环境

7.4　对比分析

高质量的数据集在离线 RL 学习中扮演着重要角色。3 个实验环境提供的数据集具有不完全相同的属性。D4RL 环境中的数据集涵盖了现实世界场景中的一系列具有挑战性的属性，包括狭窄和有偏见的数据分布、无方向和多任务的数据、不可代表行为策略、非马尔可夫行为策略、稀疏奖励、次优数据、现实领域与部分可观察性。RL Unplugged 环境中的数据集属性包括稀疏奖励、次优数据、现实领域和部分可观察性。NeoRL 环境中的数据集属性包括狭窄和有偏见的数据分布、现实领域与部分可观察性。

下面对比 D4RL 与 RL Unplugged 环境的优劣。(1)D4RL 环境中的行为策略具有非马尔可夫性质。然而，RL Unplugged 不具备这样的保证，因此通常需要其行为策略是可表示的；(2)D4RL 为了评估噪声对算法的意义，在一些环境中用随机策略、混合策略和专家策略收集不同的数据集。但 RL Unplugged 将数据限制在在线方法训练过的行为策略上，这导致大部分数据来自于混合策略和专家策略；(3)D4RL 采用了易于学习的 PyTorch 库，这有助于促进研究者之间的协作和信息交流。然而，RL Unplugged 使用 TensorFlow 库，它相对于 PyTorch 而言较为复杂，但为那些偏好使用 TensorFlow 的研究者提供了便利；(4)RL Unplugged 不仅用于测试离线 RL 算法，还包括对部分环境的在线测试，但 D4RL 仅用于测试离线 RL 算法。

与上述两类环境相比，NeoRL 的优势在于提供的数据集更接近真实世界。D4RL 和 RL Unplugged 通常存在显著的现实差距。它们是由高度探索性的

策略收集的大型数据集，并且训练的策略直接在环境中进行评估。然而，在现实世界中，探索能力相对较低，数据也非常有限，并且在评估之前应该对经过训练的策略进行充分验证。这种差异突显了在离线 RL 中面临的挑战，即将算法从仿真环境推广到现实世界，以准确地反映现实环境的特征和限制。但是由于 NeoRL 发展历程较短，且现实数据对算法的要求更为严格，所以其使用率较低。

在众多离线 RL 工作中，研究者更趋向于选择 D4RL 环境。这是因为 D4RL 不仅拥有数据集属性丰富的特点，而且提供了庞大的基线对比数据，同时其实验环境设计容易理解和复现。在离线 RL 领域进行实验和评估时，这些优势使 D4RL 成为研究者的首要选择。并且这些特性也被视为离线环境的一种选用标准，推动着离线 RL 工作的研究和发展。

下面直观地对比 3 个离线 RL 实验环境，包括 D4RL 与 NeoRL 中的所有环境以及 RL Unplugged 中的关键环境。此外，总结了每个环境所满足的问题类型、环境类型与马尔可夫性质，符号“●”表示属于该类，具体如表 7 所示。

表 7 离线强化学习 3 大实验环境对比属性

实验环境	套件	环境	控制问题		图像问题	随机性环境	确定性环境	MDP	POMDP
			离散空间	连续空间					
D4RL[127]	迷宫世界	Maze2D		●			●	●	
		AntMaze		●			●	●	
	Gym-MuJoCo[130]	HalfCheetah		●			●	●	
		Hopper		●			●	●	
		Walker2D		●			●	●	
	Adroit[131]	旋转笔		●			●	●	
		锤钉子		●			●	●	
		开门		●			●	●	
	Flow[132]			●			●	●	
	FrankaKitchen[133]			●			●	●	
	CARLA[134]			●	●		●		●
RL Unplugged[128]	DM 控制[136]	手指转动		●			●	●	
		鱼游泳		●			●	●	
		插入钉子		●		●	●	●	
	DM 运动[137]	人形差距		●	●		●	●	
		啮齿动物逃脱		●	●		●		●
		啮齿动物迷宫		●	●		●		●
	ALE[138]	太空入侵者	●		●	●		●	
		乒乓球对战	●		●	●		●	
		打砖块	●		●	●		●	
	RWRL[139]	车杆摆动		●			●		●
		步行者		●			●		●
		人形行走		●			●		●
NeoRL[129]	Gym-MuJoCo[130]	HalfCheetah		●			●	●	
		Hopper		●			●	●	
		Walker2D		●			●	●	
	工业基准[144]			●		●			●
	股票交易模拟器[145]			●		●			●
	城市管理[146]			●		●			●

8 离线强化学习应用

目前，离线 RL 方法已经被广泛应用于推荐系统、导航驾驶、自然语言处理、机器人控制以及医疗与能源等现实世界应用领域，以解决现实生活中与环境互动带来的高成本和危险性等问题。图 10 描述了从 2019 年至 2022 年（截至 2023 年 1 月 1 日），离线 RL 在不同现实世界应用领域的论文数量占比情况（共 106 篇）。图 11 展示了离线 RL 在推荐系统、导航驾驶与机器人控制领域的应用场景。

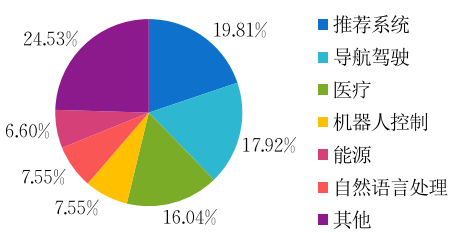


图 10 离线 RL 在现实世界应用领域的论文数量

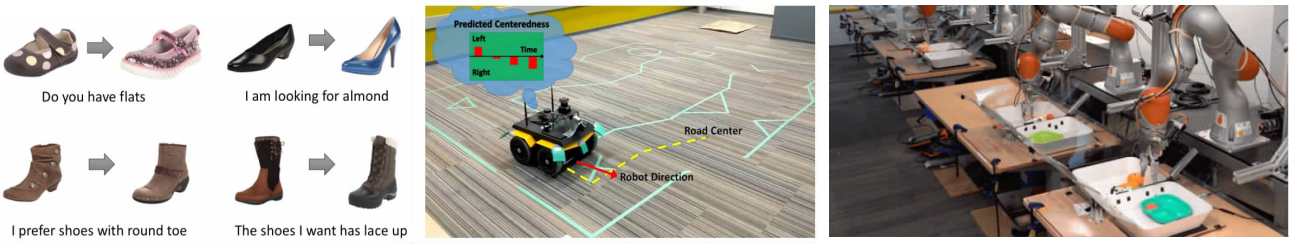


图 11　基于文本的交互式推荐、自动驾驶模拟与机器人操纵任务

8.1　推荐系统领域

推荐系统具有巨大的商业价值，RL 作为交互推荐(Interactive Recommendation)的一个强大范式，不仅使用户对系统的满意度达到最大化，而且可以不断适应用户多变的兴趣(状态)。但是，在线 RL 下的推荐系统在与现实客户互动时很容易受到限制，且成本昂贵，损害客户体验的同时要付出大量的成本。

为了解决无法与现实世界交互的问题，Zou 等人[147]提出将基于模型的离线 RL 运用到交互式推荐中，构建一个用户模拟器来模拟环境，无需与现实客户进行互动。同时，用重要性采样的方法解决了分布偏移的问题。此外，在两个大型的现实世界数据集淘宝与 Retailrocket 上验证了方法的有效性。为了多角度解决分布偏移问题，Xiao 等人[148]提出一个通用的离线交互式推荐框架，并给出 5 种正则化技术解决分布偏移问题：支持约束、监督正则化、策略约束、双重约束与奖励外推。该框架在两个公开的现实世界数据集 RecSys 和 Kaggle 上进行了广泛的实验，证明了该方法与现有的监督学习等方法相比，在推荐方面取得了卓越的性能。

具有自然语言反馈的交互式推荐系统可以提供更丰富的用户信息，并且显示出比传统推荐系统更强大的优势。Zhang 等人[30]提出一个基于文本的交互式推荐系统，从用户的自然语言反馈中提取意图，根据设备中的离线个性化数据进行策略训练。并利用 CQL[12]算法解决了分布偏移问题。此外，该推荐系统已经成功应用到智能音箱中，在亚马逊的 Echo Show 和谷歌的 Home Hub 中均得到验证。

8.2　导航驾驶领域

目前离线 RL 在生态驾驶方面取得了不错的进展。Zhu 等人[32]提出一种基于模型的安全强化学习方法用于生态驾驶。与以往无模型方法相比，该方法不需要任何外部奖励机制，简化了设计过程并提高了最终性能。同时，利用 BCQ[11]算法解决了行为策略分布与当前策略分布不匹配导致的外推误差问题。此外，该方法对于自动驾驶与机器人控制领域均有借鉴意义。

大量研究表明，许多动物在自我定位和路径规划方面形成空间表达的能力，都依赖于大脑对原始感知信号的特征编码。在自动驾驶模拟任务中，DeepDriving[149]利用人类驾驶员收集的离线数据，从图像中感知多车道驾驶的道路角度。但是，DeepDriving 通过启发式方法和规则来控制车辆，而不是利用 RL 来学习策略。基于此，Graves 等人[150]提出了一种多时间尺度(Multi-Time-Scale)预测表征学习方法，以离线方式有效地学习驾驶策略。这些策略能够很好地泛化到离线训练数据中没有涵盖的新型道路几何形状、损坏和分散注意力的车道状况。相比于 DeepDriving，该方法是对未来的长期反事实预测，而并非对当前车道中心度和道路角度的预测。

8.3　机器人控制领域

在 D4RL 实验环境 Adroit 和 FrankaKitchen 任务中，离线 RL 算法实现了对机器手臂和机器人的控制。此外，在现实世界中的机器人控制任务中，离线 RL 也取得了若干研究成果。

Lee 等人[37]从固定的离线数据集中直接学习 Q 网络。同时，为了从稀疏奖励中学习到目标 Q 函数，机器人利用事后经验回放(Hindsight Experience Replay，HER)离线地进行训练，完成了学习织物折叠的任务。然而，在现实世界中机器人只掌握一项技能无法将其作用发挥至最大。为了进行大规模多任务机器人训练，Chebotar 等人[151]对数据集中的所有轨迹和子序列进行了事后重新标记(Hindsight Relabeling)，并以完全离线的方式训练了一个目标趋向的 Q 函数。同时，利用类似的 CQL[12]算法，正则化分布外的 Q 函数。此方法能够对各种技能进行训练，包括挑选特定对象、放入各种固定装置、将货架上的物品摆放整齐、重新排列和用毛巾覆盖对象等。

9 离线强化学习展望

综上所述，离线 RL 算法给应用落地带来了巨大的价值，但现阶段仍存在一些局限性和不足之处。为了使算法更具通用性、高性能和灵活性，离线 RL 未来需要朝以下几个方向发展。

(1) 异策略评估问题

异策略评估(Off-Policy Evaluation，OPE)是指仅通过经验对策略进行评估，如果在相同静态数据集上一直进行训练很可能导致过拟合现象的出现。所以一个好的 OPE 方法对离线 RL 至关重要，目前先进的 OPE 方法有重要性采样、基于模型和拟合 Q 评估(FQE)等，但是仍出现评估不准确以及无法在庞大数据集上有良好的效果等问题。因此，寻找一个能够在评估阶段验证策略好坏的 OPE 方法，可以极大地提升离线 RL 算法的性能。同时在实践中，大多数离线 RL 不依赖于 OPE 方法评估性能，而是在固定数量的步骤中使用一组超参数进行训练，并通过上次迭代的策略来在线评估其质量。这种方法不仅需要花费大量的时间和计算资源来寻找一个最优的超参数组合，还常导致智能体只能得到次优的策略。总之，对于离线 RL 问题而言，异策略评估与超参数调整仍是一个值得研究的开放性问题。

(2) 有效权衡策略的累积误差

在 RL 问题中，状态-动作空间往往是十分庞大的，使得完全存储和表达 Q 函数是不可行的。因此，通过神经网络来近似 Q 函数，但函数近似带来的误差能够对学习产生不利影响，从而导致对 Q 值的过高估计。在在线 RL 中，函数近似可以通过主动收集数据来修正高估的误差。然而离线 RL 中，这些误差将逐渐累积并影响到后续迭代的过程。目前，已经通过正则化、不确定性度量和生成模型等技术来解决该问题。在未来展望中，应当有效地权衡当前学习策略的累积误差和次优性，并且可以在实践中通过标准化技术实现，而无须额外的函数近似器来拟合行为策略。

(3) 扩展离线 RL 方法的研究内容

由于基于无模型的离线 RL 算法发展历程较长、易于实现且资源消耗相对较少，因此目前离线学习领域的研究主要集中在无模型方法。与此同时，基于模型和基于 Transformer 模型的算法也在迅速发展。将离线 RL 方法运用到其他技术中，可以提供新的解决思路。目前，离线 RL 已经与元学习[55,152-153]、分层学习[154-155]、联邦学习[156-157]和分布式学习[158]等方法相结合。其中，元学习可以使 RL 算法快速地适应不可见的任务。分层学习运用时序抽象表达方式为离线 RL 提供施加结构。联邦学习使 RL 在保护隐私的环境中更为有效，并充分利用分散的和多样化的数据。分布式学习得到累计奖励的分布而非期望值，可以在风险中与风险敏感的领域上表现出强大性能。总体而言，受到实际问题的驱动，离线 RL 的研究领域正在不断扩展。从无模型到基于模型再到基于 Transformer 模型的方法，旨在寻求更灵活和高效的解决方案。在未来，将离线 RL 与其他技术相结合为其应用在实用场景中提供更多可能性。

(4) 符合现实世界要求的实验环境

离线 RL 的突出特点是对现实世界的适用性，从而减少不必要的风险以及节约人力和物力的成本。然而，现实世界的数据往往是非常复杂和难以获取的，因此大多数研究和实验环境都被限制在模拟环境中，例如 D4RL。尽管人们已经在努力设计能够捕获现实世界属性的实验环境，例如 NeoRL。该环境提供了工业、金融和管理上的真实数据集，但是依旧没有涉足导航、定位和无人驾驶等方面。总体来看，设计一个现实世界的离线 RL 实验环境仍是一项重要问题，其应包含多个场景，每个场景都有不同的难度与复杂度，同时涵盖不同的任务类型和环境属性。

10 结束语

近些年，离线 RL 方法的研究热度不断增加，并在短时间内出现了数百篇高质量的文章。现有 RL 技术通常基于静态环境的理论假设进行开发与测试，而现实应用场景打破了现有 RL 理论的假设条件。因此，基于离线数据寻找最优策略，试图突破静态环境假设的方法变得尤为重要。

离线 RL 的一个重要问题是分布偏移，用来生成数据集的行为策略分布与当前学习策略的分布不一致。因此，离线 RL 必须设计一种能在给定的行为分布范围内采取动作、并在其之外获得良好性能的算法。为了解决该问题，从无模型、基于模型与基于 Transformer 模型 3 个方向出发，详细描述了离线 RL 方法的研究现状，对比并分析了各个算法的特点。尽管 3 个分类有着完全不同的核心思想和求解技术，但是其共性是十分明显的。一方面，这些方

法几乎都采用了 MDP 理论和离线数据，都希望实现更安全、更高效的现实 RL 应用落地。但是，3 个分类也存在不同之处，无模型与基于模型的离线 RL 方法旨在解决分布偏移与外推误差的问题，而基于 Transformer 模型的离线 RL 方法可以同时对环境和行为策略进行建模，无需担心离线 RL 中的外推误差问题。该方法更加关注的是长时序与稀疏奖励这类经典 RL 问题。同时，为了对离线 RL 性能评估提供统一的标准，研究者给出了多种任务的数据集来测试不同算法的效果。目前，离线 RL 较为流行的 3 类实验环境为 D4RL、RL Unplugged 与 NeoRL。

虽然离线 RL 已成功地应用在推荐系统、导航驾驶、机器人控制和自然语言处理等现实领域。但在整体上，其研究进展还处在初步阶段，未来仍需要朝以下几个方向发展：(1) 选择合适的异策略评估方法；(2) 有效地权衡学习策略的累积误差，不局限于使用函数近似等方法；(3) 将元学习、分层学习、联邦学习和分布式学习等方法与离线 RL 相结合，提高学习效率和性能；(4) 设计一个更加符合现实世界的实验环境。可以预见的是，离线 RL 一旦取得突破性进展，RL 的自主决策能力将极大地推动产业变革，并且为人类社会带来巨大的价值。

参　考　文　献

[1] Sutton R S, Barto A G. Reinforcement Learning: An Introduction. Cambridge, USA: MIT Press, 2018

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

[3] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436-444

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

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

[6] Ernst D, Geurts P, Wehenkel L. Tree-based batch mode reinforcement learning. Journal of Machine Learning Research, 2005, 6(1): 503-556

[7] Riedmiller M. Neural fitted Q iteration — first experiences with a data efficient neural reinforcement learning method// Proceedings of the European Conference on Machine Learning. Porto, Portugal, 2005: 317-328

[8] Lange S, Gabel T, Riedmiller M. Batch reinforcement learning. Reinforcement Learning: State-of-the-Art, 2012, 12(1): 45-73

[9] Levine S, Kumar A, Tucker G, et al. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv preprint arXiv:2005.01643, 2020

[10] Prudencio R F, Maximo M R, Colombini E L. A survey on offline reinforcement learning: Taxonomy, review, and open problems. IEEE Transactions on Neural Networks Learning Systems, 2023, 99(1): 1-21

[11] Fujimoto S, Meger D, Precup D. Off-policy deep reinforcement learning without exploration//Proceedings of the International Conference on Machine Learning. Los Angeles, USA, 2019: 2052-2062

[12] Kumar A, Zhou A, Tucker G, et al. Conservative Q-learning for offline reinforcement learning//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020: 1179-1191

[13] Chen X, Zhou Z, Wang Z, et al. BAIL: Best-action imitation learning for batch deep reinforcement learning//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020: 18353-18363

[14] Siegel N Y, Springenberg J T, Berkenkamp F, et al. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. arXiv preprint arXiv: 2002.08396, 2020

[15] Wang Z, Novikov A, Zolna K, et al. Critic regularized regression//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020: 7768-7778

[16] Emmons S, Eysenbach B, Kostrikov I, et al. RvS: What is essential for offline RL via supervised learning?//Proceedings of the International Conference on Learning Representations. Virtual, 2022: 1-14

[17] Xu H, Jiang L, Li J, et al. A policy-guided imitation approach for offline reinforcement learning//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022: 4085-4098

[18] Yu T, Thomas G, Yu L, et al. MOPO: Model-based offline policy optimization//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020: 14129-14142

[19] Kidambi R, Rajeswaran A, Netrapalli P, et al. MORel: Model-based offline reinforcement learning//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020: 21810-21823

[20] Rafailov R, Yu T, Rajeswaran A, et al. Offline reinforcement learning from images with latent space models//Proceedings of the Learning for Dynamics and Control. Virtual, 2021: 1154-1168

[21] Yang Y, Jiang J, Zhou T, et al. Pareto policy pool for model-based offline reinforcement learning//Proceedings of the International Conference on Learning Representations. Virtual, 2022: 1-22

[22] Matsushima T, Furuta H, Matsuo Y, et al. Deployment-efficient reinforcement learning via model-based offline optimization//Proceedings of the International Conference on Learning Representations. Vienna, Austria, 2021: 1-21

[23] Yu T, Kumar A, Rafailov R, et al. COMBO: Conservative offline model-based policy optimization//Proceedings of the Advances in Neural Information Processing Systems. Los Angeles, USA, 2021: 28954-28967

[24] Chen L, Lu K, Rajeswaran A, et al. Decision transformer: Reinforcement learning via sequence modeling//Proceedings of the Advances in Neural Information Processing Systems. Los Angeles, USA, 2021: 15084-15097

[25] Janner M, Li Q, Levine S. Offline reinforcement learning as one big sequence modeling problem//Proceedings of the Advances in Neural Information Processing Systems. Los Angeles, USA, 2021: 1273-1286

[26] Zheng Q, Zhang A, Grover A. Online decision transformer//Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 27042-27059

[27] Furuta H, Matsuo Y, Gu S S. Generalized decision transformer for offline hindsight information matching//Proceedings of the International Conference on Learning Representations. Virtual, 2022: 1-28

[28] Shang J, Kahatapitiya X L K, Lee Y-C, et al. StARformer: Transformer with state-action-reward representations for robot learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(11): 12862-12877

[29] Gilotte A, Calauzènes C, Nedelec T, et al. Offline A/B testing for recommender systems//Proceedings of the Web Search and Data Mining. Marina Del Rey, USA, 2018: 198-206

[30] Zhang R, Yu T, Shen Y, et al. Text-based interactive recommendation via offline reinforcement learning//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2022: 11694-11702

[31] Deffayet R, Thonet T, Renders J-M, et al. Offline evaluation for reinforcement learning-based recommendation: A critical issue and some alternatives. arXiv preprint arXiv: 2301.00993, 2023

[32] Zhu Z, Pivaro N, Gupta S, et al. Safe model-based off-policy reinforcement learning for eco-driving in connected and automated hybrid electric vehicles. IEEE Transactions on Intelligent Vehicles, 2022, 7(2): 387-398

[33] Diehl C, Sievernich T, Krüger M, et al. UMBRELLA: Uncertainty-aware model-based offline reinforcement learning leveraging planning. arXiv preprint arXiv:2111.11097, 2021

[34] Diehl C, Sievernich T S, Krüger M, et al. Uncertainty-aware model-based offline reinforcement learning for automated driving. IEEE Robotics and Automation Letters, 2023, 8(2): 1167-1174

[35] Jaques N, Shen J H, Ghandeharioun A, et al. Human-centric dialog training via offline reinforcement learning//Proceedings of the Empirical Methods in Natural Language Processing. Virtual, 2020: 3985-4003

[36] Jang Y, Lee J, Kim K-E. GPT-Critic: Offline reinforcement learning for end-to-end task-oriented dialogue systems//Proceedings of the International Conference on Learning Representations. Vienna, Austria, 2021: 1-16

[37] Lee R, Ward D, Cosgun A, et al. Learning arbitrary-goal fabric folding with one hour of real robot experience//Proceedings of the Conference on Robot Learning. Cambridge, USA, 2020: 2317-2327

[38] Schmeckpeper K, Rybkin O, Daniilidis K, et al. Reinforcement learning with videos: Combining offline observations with interaction//Proceedings of the Conference on Robot Learning. Cambridge, USA, 2020: 339-354

[39] Shiranthika C, Chen K-W, Wang C-Y, et al. Supervised optimal chemotherapy regimen based on offline reinforcement learning. IEEE Journal of Biomedical and Health Informatics, 2022, 26(9): 4763-4772

[40] Killian T W, Zhang H, Subramanian J, et al. An empirical study of representation learning for reinforcement learning in healthcare//Proceedings of the Machine Learning for Health Workshop. Virtual, 2020: 139-160

[41] Zhan X, Xu H, Zhang Y, et al. DeepThermal: Combustion optimization for thermal power generating units using offline reinforcement learning//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2022: 4680-4688

[42] Jang D, Spangher L, Srivistava T, et al. Offline-online reinforcement learning for energy pricing in office demand response: Lowering energy and data costs//Proceedings of the Systems for Energy-Efficient Buildings, Cities, and Transportation. Coimbra, Portugal, 2021: 131-139

[43] Zhang G, Zhang C, Wang W, et al. Offline reinforcement learning control for electricity and heat coordination in a supercritical CHP unit. Energy, 2023, 266(1): 126485

[44] Kaelbling L P, Littman M L, Cassandra A R. Planning and acting in partially observable stochastic domains. Artificial Intelligence, 1998, 101(1-2): 99-134

[45] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529-533

[46] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning//Proceedings of the International Conference on Learning Representations. San Juan, Puerto Rico, 2016: 1-14

[47] Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 1587-1596

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

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

[50] Haarnoja T, Zhou A, Abbeel P, et al. Soft Actor-Critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 1861-1870

[51] Konda V R, Tsitsiklis J N. Actor-critic algorithms//Proceedings of the Advances in Neural Information Processing Systems. Cambridge, USA, 1999: 1008-1014

[52] Fujimoto S, Conti E, Ghavamzadeh M, et al. Benchmarking batch deep reinforcement learning algorithms. arXiv preprint arXiv:1910.01708, 2019

[53] Formanek C, Jeewa A, Shock J, et al. Off-the-Grid MARL: A framework for dataset generation with baselines for cooperative offline multi-agent reinforcement learning. arXiv preprint arXiv:2302.00521, 2023

[54] Jin J, Graves D, Haigh C, et al. Offline learning of counterfactual predictions for real-world robotic reinforcement learning //Proceedings of the IEEE International Conference on Robotics and Automation. Philadelphia, USA, 2022: 3616-3623

[55] Li L, Yang R, Luo D. FOCAL: Efficient fully-offline meta-reinforcement learning via distance metric learning and behavior regularization//Proceedings of the International Conference on Learning Representations. Virtual, 2020: 1-23

[56] Cheng C-A, Xie T, Jiang N, et al. Adversarially trained actor critic for offline reinforcement learning//Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 3852-3878

[57] Luo F-M, Xu T, Lai H, et al. A survey on model-based reinforcement learning. arXiv preprint arXiv: 2206.09328, 2022

[58] 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): 1-25(in Chinese)
(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)：1-25)

[59] Kingma D P, Welling M. Auto-encoding variational Bayes//Proceedings of the International Conference on Learning Representations. Banff, Canada, 2014: 1-14

[60] Kumar A, Fu J, Soh M, et al. Stabilizing off-policy Q-learning via bootstrapping error reduction//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2019: 11761-11771

[61] Wu J, Wu H, Qiu Z, et al. Supported policy optimization for offline reinforcement learning//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022: 31278-31291

[62] Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2015: 3483-3491

[63] Zhou W, Bajracharya S, Held D. PLAS: Latent action space for offline reinforcement learning//Proceedings of the Conference on Robot Learning. Cambridge, USA, 2021: 1719-1735

[64] Chen X, Ghadirzadeh A, Yu T, et al. Latent-variable advantage-weighted policy optimization for offline RL. arXiv preprint arXiv:2203.08949, 2022

[65] Gu P, Zhao M, Chen C, et al. Learning pseudometric-based action representations for offline reinforcement learning// Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 7902-7918

[66] Lyu J, Ma X, Li X, et al. Mildly conservative Q-learning for offline reinforcement learning//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022: 1711-1724

[67] Lou X, Yin Q, Zhang J, et al. Offline reinforcement learning with representations for actions. Information Sciences, 2022, 610(1): 746-758

[68] Singh A, Liu H, Zhou G, et al. Parrot: Data-driven behavioral priors for reinforcement learning//Proceedings of the International Conference on Learning Representations. Virtual, 2020: 1-19

[69] Akimov D, Kurenkov V, Nikulin A, et al. Let offline RL flow: Training conservative agents in the latent space of normalizing flows. arXiv preprint arXiv:2211.11096, 2022

[70] Wang Z, Hunt J J, Zhou M. Diffusion policies as an expressive policy class for offline reinforcement learning//Proceedings of the International Conference on Learning Representations. Kigali, Rwanda, 2023: 1-16

[71] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020: 6840-6851

[72] Ray A, Achiam J, Amodei D. Benchmarking safe exploration in deep reinforcement learning. arXiv preprint arXiv: 1910.01708, 2019

[73] Dong W, Liu S, Sun S. Safe batch constrained deep reinforcement learning with generative adversarial network. Information Sciences, 2023, 634(1): 259-270

[74] Weissenbacher M, Sinha S, Garg A, et al. Koopman Q-learning: Offline reinforcement learning via symmetries of dynamics//Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 23645-23667

[75] Yu T, Quillen D, He Z, et al. Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning //Proceedings of the Conference on Robot Learning. Osaka, Japan, 2020: 1094-1100

[76] Zhu Y, Wong J, Mandlekar A, et al. Robosuite: A modular simulation framework and benchmark for robot learning. arXiv preprint arXiv:2009.12293, 2020

[77] Mazoure B, Kostrikov I, Nachum O, et al. Improving zero-shot generalization in offline reinforcement learning using generalized similarity functions//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022: 25088-25101

[78] Cobbe K, Hesse C, Hilton J, et al. Leveraging procedural generation to benchmark reinforcement learning//Proceedings of the International Conference on Machine Learning. Virtual, 2020: 2048-2056

[79] Zang H, Li X, Yu J, et al. Behavior prior representation learning for offline reinforcement learning//Proceedings of the International Conference on Learning Representations. Kigali, Rwanda, 2023: 1-34

[80] Pomerleau D A. ALVINN: An autonomous land vehicle in a neural network//Proceedings of the Advances in Neural Information Processing Systems. Denver, USA, 1988: 305-313

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

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

[83] Ho J, Ermon S. Generative adversarial imitation learning//Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016: 4565-4573

[84] Abdolmaleki A, Springenberg J T, Tassa Y, et al. Maximum a posteriori policy optimisation//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018: 1-23

[85] Vinyals O, Babuschkin I, Czarnecki W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature, 2019, 575(7782): 350-354

[86] Kumar A, Hong J, Singh A, et al. Should I run offline reinforcement learning or behavioral cloning?//Proceedings of the International Conference on Learning Representations. Virtual, 2022: 1-36

[87] Sutton R S, Mcallester D, Singh S, et al. Policy gradient methods for reinforcement learning with function approximation //Proceedings of the Advances in Neural Information Processing Systems. Cambridge, USA, 1999: 1057-1063

[88] Agarwal R, Schuurmans D, Norouzi M. An optimistic perspective on offline reinforcement learning//Proceedings of the International Conference on Machine Learning. Virtual, 2020: 104-114

[89] Wu Y, Tucker G, Nachum O. Behavior regularized offline reinforcement learning. arXiv preprint arXiv: 1911.11361, 2019

[90] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of Wasserstein GANs//Proceedings of the Advances in Neural Information Processing Systems. Los Angeles, USA, 2017: 5767-5777

[91] Fujimoto S, Gu S S. A minimalist approach to offline reinforcement learning//Proceedings of the Advances in Neural Information Processing Systems. Los Angeles, USA, 2021: 20132-20145

[92] Ran Y, Li Y, Zhang F, et al. Policy regularization with dataset constraint for offline reinforcement learning//Proceedings of the International Conference on Machine Learning. Honolulu, USA, 2023: 28701-28717

[93] Peters J, Schaal S. Reinforcement learning by reward-weighted regression for operational space control//Proceedings of the International Conference on Machine Learning. Corvallis, USA, 2007: 745-750

[94] Peng X B, Kumar A, Zhang G, et al. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. arXiv preprint arXiv:1910.00177, 2019

[95] Nair A, Gupta A, Dalal M, et al. AWAC: Accelerating online reinforcement learning with offline datasets. arXiv preprint arXiv:2006.09359, 2020

[96] Gulcehre C, Colmenarejo S G, Wang Z, et al. Regularized behavior value estimation. arXiv preprint arXiv:2103.09575, 2021

[97] Brandfonbrener D, Whitney W, Ranganath R, et al. Offline RL without off-policy evaluation//Proceedings of the Advances in Neural Information Processing Systems. Los Angeles, USA, 2021: 4933-4946

[98] Kostrikov I, Fergus R, Tompson J, et al. Offline reinforcement learning with Fisher divergence critic regularization//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 5774-5783

[99] Johnson O. Information Theory and the Central Limit Theorem. London, UK: World Scientific, 2004

[100] Kostrikov I, Nair A, Levine S. Offline reinforcement learning with implicit Q-learning//Proceedings of the Conference on Learning Representations. Virtual, 2022: 1-13

[101] Wu Y, Zhai S, Srivastava N, et al. Uncertainty weighted actor-critic for offline reinforcement learning//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 11319-11328

[102] An G, Moon S, Kim J-H, et al. Uncertainty-based offline reinforcement learning with diversified Q-ensemble//Proceedings of the Advances in Neural Information Processing Systems. Los Angeles, USA, 2021: 7436-7447

[103] Bai C, Wang L, Yang Z, et al. Pessimistic bootstrapping for uncertainty-driven offline reinforcement learning//Proceedings of the International Conference on Learning Representations. Virtual, 2022: 1-29

[104] Yang R, Bai C, Ma X, et al. RORL: Robust offline reinforcement learning via conservative smoothing//Proceedings of the Advances in Neural Information Processing Systems. New Orleans, USA, 2022: 23851-23866

[105] Sutton R S, Precup D, Singh S. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. Artificial Intelligence, 1999, 112(1-2): 181-211

[106] Peng X B, Abbeel P, Levine S, et al. DeepMimic: Example-guided deep reinforcement learning of physics-based character skills. ACM Transactions on Graphics, 2018, 37(4): 1-14

[107] Ahn M, Zhu H, Hartikainen K, et al. ROBEL: Robotics benchmarks for learning with low-cost robots//Proceedings of the Conference on Robot Learning. Osaka, Japan, 2020: 1300-1313

[108] Yu T, Kumar A, Chebotar Y, et al. How to leverage unlabeled data in offline reinforcement learning//Proceedings of the International Conference on Machine Learning. Baltimore, USA, 2022: 25611-25635

[109] Hafner D, Lillicrap T, Fischer I, et al. Learning latent dynamics for planning from pixels//Proceedings of the International Conference on Machine Learning. Los Angeles, USA, 2019: 2555-2565

[110] Zhang M, Vikram S, Smith L, et al. SOLAR: Deep structured representations for model-based reinforcement learning//Proceedings of the International Conference on Machine Learning. Los Angeles, USA, 2019: 7444-7453

[111] Yu Yang. Offline data reinforcement learning: Approaches and progress. China Basic Science, 2022, 24(3): 35-39+46 (in Chinese)
(俞扬. 离线数据强化学习：途径与进展. 中国基础科学, 2022, 24(3): 35-39+46)

[112] Deisenroth M, Rasmussen C E. PILCO: A model-based and data-efficient approach to policy search//Proceedings of the International Conference on Machine Learning. Virtual, 2011: 465-472

[113] Chua K, Calandra R, Mcallister R, et al. Deep reinforcement learning in a handful of trials using probabilistic dynamics models//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2018: 4759-4770

[114] Sutton R S. Dyna, an integrated architecture for learning, planning, and reacting. ACM Sigart Bulletin, 1991, 2(4): 160-163

[115] Luo Y, Xu H, Li Y, et al. Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019: 1-28

[116] Janner M, Fu J, Zhang M, et al. When to trust your model: Model-based policy optimization//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2019: 12498-12509

[117] Tassa Y, Erez T, Todorov E. Synthesis and stabilization of complex behaviors through online trajectory optimization//Proceedings of the International Conference on Intelligent Robots and Systems. Vilamoura, Portugal, 2012: 4906-4913

[118] Feinberg V, Wan A, Stoica I, et al. Model-based value estimation for efficient model-free reinforcement learning. arXiv preprint arXiv:1803.00101, 2018

[119] Trabucco B, Kumar A, Geng X, et al. Conservative objective models for effective offline model-based optimization//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 10358-10368

[120] Azar M G, Osband I, Munos R. Minimax regret bounds for reinforcement learning//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 263-272

[121] Kostrikov I, Agrawal K K, Dwibedi D, et al. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019: 1-14

[122] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the Advances in Neural Information Processing Systems. Los Angeles, USA 2017: 5998-6008

[123] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020: 1877-1901

[124] Ramesh A, Pavlov M, Goh G, et al. Zero-shot text-to-image generation//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 8821-8831

[125] Mesnard T, Weber T, Viola F, et al. Counterfactual credit assignment in model-free reinforcement learning//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 7654-7664

[126] Liu M, Zhu M, Zhang W. Goal-conditioned reinforcement learning: Problems and solutions//Proceedings of the International Joint Conference on Artificial Intelligence. Vienna, Austria, 2022: 5502-5511

[127] Fu J, Kumar A, Nachum O, et al. D4RL: Datasets for deep data-driven reinforcement learning. arXiv preprint arXiv: 2004.07219, 2020

[128] Gulcehre C, Wang Z, Novikov A, et al. RL unplugged: A suite of benchmarks for offline reinforcement learning//Proceedings of the Advances in Neural Information Processing Systems. Virtual, 2020: 7248-7259

[129] Qin R, Gao S, Zhang X, et al. NeoRL: A near real-world benchmark for offline reinforcement learning. arXiv preprint arXiv:2102.00714, 2021

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

[131] Rajeswaran A, Kumar V, Gupta A, et al. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations//Proceedings of the Robotics: Science and Systems. Pittsburgh, USA, 2018: 1-9

[132] Wu C, Kreidieh A, Parvate K, et al. Flow: Architecture and benchmarking for reinforcement learning in traffic control. arXiv preprint arXiv:1710.05465, 2017

[133] Gupta A, Kumar V, Lynch C, et al. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning//Proceedings of the Conference on Robot Learning. Osaka, Japan, 2019: 1025-1037

[134] Dosovitskiy A, Ros G, Codevilla F, et al. CARLA: An open urban driving simulator//Proceedings of the Conference on Robot Learning. Mountain View, USA, 2017: 1-16

[135] Brockman G, Cheung V, Pettersson L, et al. OpenAI Gym. arXiv preprint arXiv:1606.01540, 2016

[136] Tassa Y, Doron Y, Muldal A, et al. DeepMind control suite. arXiv preprint arXiv:1801.00690, 2018

[137] Tunyasuvunakool S, Muldal A, Doron Y, et al. Dm_control: Software and tasks for continuous control. Software Impacts, 2020, 6(1): 100022

[138] Bellemare M G, Naddaf Y, Veness J, et al. The arcade learning environment: An evaluation platform for general agents. Journal of Artificial Intelligence Research, 2013, 47(1): 253-279

[139] Dulac-Arnold G, Levine N, Mankowitz D J, et al. An empirical investigation of the challenges of real-world reinforcement learning. arXiv preprint arXiv:2003.11881, 2020

[140] Hasselt H V, Guez A, Silver D. Deep reinforcement learning with double Q-learning//Proceedings of the AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016: 2094-2100

[141] Kaiser L, Babaeizadeh M, Milos P, et al. Model-based reinforcement learning for Atari//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 1-28

[142] Reddy S, Dragan A D, Levine S. SQIL: Imitation learning via reinforcement learning with sparse rewards//Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 1-14

[143] Huang Zhi-Gang, Liu Quan, Zhang Li-Hua, et al. A survey of inverse reinforcement learning. Journal of Software, 2023, 34(2): 733-760(in Chinese)
(黄志刚, 刘全, 张立华等. 深度分层强化学习研究与发展. 软件学报, 2023, 34(2): 733-760)

[144] Hein D, Depeweg S, Tokic M, et al. A benchmark environment motivated by industrial control problems//Proceedings of the IEEE Symposium Series on Computational Intelligence. Honolulu, USA, 2017: 1-8

[145] Liu X-Y, Yang H, Chen Q, et al. FinRL: A deep reinforcement learning library for automated stock trading in quantitative finance. arXiv preprint arXiv:2011.09607, 2020

[146] Vázquez-Canteli J R, Kämpf J, Henze G, et al. CityLearn v1.0: An OpenAI Gym environment for demand response with deep reinforcement learning//Proceedings of the Systems for Energy-Efficient Buildings, Cities, and Transportation. New York, USA, 2019: 356-357

[147] Zou L, Xia L, Du P, et al. Pseudo Dyna-Q: A reinforcement learning framework for interactive recommendation//Proceedings of the International Conference on Web Search and Data Mining. Houston, USA, 2020: 816-824

[148] Xiao T, Wang D. A general offline reinforcement learning framework for interactive recommendation//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2021: 4512-4520

[149] Chen C, Seff A, Kornhauser A, et al. DeepDriving: Learning affordance for direct perception in autonomous driving//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 2722-2730

[150] Graves D, Nguyen N M, Hassanzadeh K, et al. Learning robust driving policies without online exploration//Proceedings of the IEEE International Conference on Robotics and Automation. Cambridge, USA, 2021: 13186-13193

[151] Chebotar Y, Hausman K, Lu Y, et al. Actionable models: Unsupervised offline reinforcement learning of robotic skills//Proceedings of the International Conference on Machine Learning. Virtual, 2021: 1518-1528

[152] Dorfman R, Shenfeld I, Tamar A. Offline meta reinforcement learning—Identifiability challenges and effective data collection strategies//Proceedings of the Advances in Neural Information Processing Systems. Los Angeles, USA, 2021: 4607-4618

[153] Zhou R, Gao C, Zhang Z, et al. Generalizable task representation learning for offline meta-reinforcement learning with data limitations//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 17132-17140

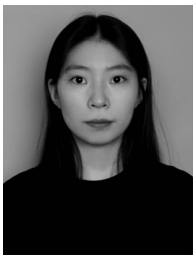
[154] Lu K, Grover A, Abbeel P, et al. Reset-free lifelong learning with skill-space planning//Proceedings of the International Conference on Learning Representations. Vienna, Austria, 2021: 1-20

[155] Park S, Ghosh D, Eysenbach B, et al. HIQL: Offline goal-conditioned RL with latent states as actions//Proceedings of the Advances in Neural Information Processing Systems. Orleans, USA, 2023: 1-26

[156] Xie Z, Song S. FedKL: Tackling data heterogeneity in federated reinforcement learning by penalizing KL divergence. IEEE Journal on Selected Areas in Communications, 2023, 41(4): 1227-1242

[157] Rengarajan D, Ragothaman N, Kalathil D, et al. Federated ensemble-directed offline reinforcement learning. arXiv preprint arXiv:2305.03097, 2023

[158] Ma Y, Jayaraman D, Bastani O. Conservative offline distributional reinforcement learning//Proceedings of the Advances in Neural Information Processing Systems. Los Angeles, USA, 2021: 19235-19247



WU Lan, Ph. D. candidate. Her main research interests include deep reinforcement learning and offline reinforcement learning.

LIU Quan, Ph. D., professor, Ph. D. supervisor. His research interests include deep reinforcement learning and automated reasoning.

HUANG Zhi-Gang, Ph. D. candidate. His research interests include deep reinforcement learning and hierarchical reinforcement learning.

ZHANG Li-Hua, Ph. D. candidate. His research interests include deep reinforcement learning and inverse reinforcement learning.

Background

Offline reinforcement learning holds significant importance in the field of deep reinforcement learning. It involves generating static datasets from behavior policies without the need for online interactions with the environment, effectively transforming large-scale datasets into powerful decision engines. The rise of offline reinforcement learning has not only accelerated the development of decision engines but also provided researchers with a stable and efficient training framework. In recent years, offline reinforcement learning methods have garnered widespread academic attention and in-depth research, achieving notable successes in practical applications. Currently, this approach is widely applied across diverse domains, encompassing recommendation systems, navigation and autonomous driving, natural language processing, robot control, healthcare, and the energy sectors. Its application signifies a pivotal advancement in addressing real-world challenges in reinforcement learning. This technology is recognized for its potential to effectively tackle complex problems and is viewed as a promising avenue for advancing practical applications of reinforcement learning in real-world scenarios.

We have presented an exposition on the foundational underpinnings and theoretical framework of offline reinforcement learning. Furthermore, we have classified offline reinforcement learning methodologies into three principal categories, namely model-free, model-based, and transformer-based. Specifically, these methods do not share the same focus and aim to address distinct challenges, achieving incremental improvements in handling distribution shifts. Model-free offline reinforcement learning methods focus on policy evaluation and improvement by directly utilizing trajectory information from static data. In contrast, model-based offline reinforcement learning methods aim to learn dynamic environment models from static datasets to optimize policies. Recently, transformer-based offline reinforcement learning methods have attracted prominence due to their superior sequence modeling abilities, showing exceptional performance in managing complex environments and long-term sequential data. Subsequently, we have undertaken a comprehensive examination of the present research landscape and the anticipated trajectories of development for each of these delineated categories.

This paper was supported by the National Natural Science Foundation of China (Nos. 62376179, 62176175), the National Natural Science Foundation of Xinjiang Uygur Autonomous Region (No. 2022D01A238), and A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). These projects aim to enrich reinforcement learning theory and develop efficient algorithms to significantly enhance their computational power and applicability across diverse domains.