强化学习离线策略评估研究综述

王硕汝 牛温佳 童恩栋 陈 彤 李 赫 田蕴哲
刘吉强 韩 臻 李浥东

(北京交通大学智能交通数据安全与隐私保护北京市重点实验室 北京 100044)

摘 要 在强化学习应用中,为避免意外风险,需要在强化学习实际部署前进行离线策略评估(Off-Policy Evaluation, OPE),这在机器人、自动驾驶等领域产生了巨大的应用前景.离线策略评估是从行为策略收集到的轨迹数据中,不需要通过实际的强化学习而估计目标策略的状态价值,通常情况下学习目标是使所估计的目标策略状态价值与目标策略真实执行的状态价值均方误差尽可能小.行为策略与目标策略间的差异性,以及新应用中出现的行为策略奖励稀疏性,不断给离线策略评估带来了挑战.本文系统性地梳理了近二十年离线策略评估的主要方法:纯模型法、重要性采样法、混合模型法和 PU 学习法(Positive Unlabeled, PU),主要内容包括:(1)描述了离线策略评估的相关理论背景知识;(2)分别阐述了各类方法的机理、方法中模型的细节差异;(3)详细对各类方法及模型进行了机理对比,并通过实验进行了主流离线策略评估模型的程序复现与性能对比.最后展望了离线策略评估的技术挑战与可能发展方向.

关键词 人工智能;强化学习;离线策略评估;重要性采样;PU 学习
中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2022.01926

Research on Off-Policy Evaluation in Reinforcement Learning: A Survey

WANG Shuo-Ru NIU Wen-Jia TONG En-Dong CHEN Tong LI He TIAN Yun-Zhe
LIU Ji-Qiang HAN Zhen LI Yi-Dong

(*Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing Jiaotong University, Beijing* 100044)

Abstract Among reinforcement learning (RL) applications, the off-policy evaluation (OPE) is employed to avoid unexpected risks before the actual deployment, which has been applied in many fields, including the robot, autonomous driving, and so on. The core learning goal of OPE is to minimize the mean squared error of state values between the new estimate and executed values from the target policy. With the advantage of OPE, researchers can estimate target policy's state values through collected history trajectories before RL execution, further realize an evaluation of reinforcement learning policy and perform necessary policy control in advance. In recent years, on the one hand, OPE has drawn many interests from both researchers and engineers. On the other hand, due to the difference between behavior policy and target policy and possible reward sparseness of behavior policy, OPE also faces emerging challenges to overcome. This paper systematically summarizes the state-of-the-art OPE methods in latest twenty years. These methods mainly fall into four categories: directed-model based, importance-sampling based, hybrid-model based, and

收稿日期:2020-12-14;在线发布日期:2021-07-28.本课题得到国家自然科学基金(61972025,61802389,61672092,U1811264,61966009)、国家重点研发计划(2020YFB1005604,2020YFB2103802)资助.**王硕汝**,硕士研究生,主要研究方向为网络空间安全.E-mail:18120481@bjtu.edu.cn.**牛温佳**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为人工智能安全、内容安全.E-mail:niuwj@bjtu.edu.cn.**童恩栋**(通信作者),博士,讲师,主要研究方向为人工智能安全、网络安全.E-mail:edtong@bjtu.edu.cn.**陈 彤**,博士研究生,主要研究方向为网络空间安全.**李 赫**,硕士研究生,主要研究方向为网络空间安全.**田蕴哲**,硕士研究生,主要研究方向为网络空间安全.**刘吉强**,博士,教授,主要研究领域为可信计算、隐私保护、云计算安全.**韩 臻**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为计算机安全、人工智能及应用、系统安全.**李浥东**,博士,教授,中国计算机学会(CCF)杰出会员,主要研究领域为隐私保护数据分析.

PU-learning based methods. For directed-model-based methods, we describe value iteration estimator and direct model (DM). Among importance-sampling based methods, we present Importance Sampling (IS), Step Importance Sampling (step-IS), Weighted Importance Sampling (WIS), Step Weighted Importance Sampling (WIS), Regression Importance Sampling (RIS), Marginalized Importance Sampling (MIS), Incremental Importance Sampling (INCRIS). We compare hybrid-model based methods, which combines the directed-model based with the importance-sampling based methods, including Doubly Robust (DR), Weighted Doubly Robust (WDR), Model and Guided Importance Sampling Combined (MAGIC), Longitudinal Targeted Maximum Likelihood Estimator (LTMLE), TMLE for RL (RLTMLE). We also introduce the PU-learning-based method involving Off-policy Classification (OPC), Soft Off-policy Classification (SoftOPC). In addition to present fundamental theoretical knowledge of OPE, we focus on analyzing different OPE methods' differences in detail and discussing essential OPE mechanisms. We also implement the mainstream OPE models and compare their performance to provide more concrete results for better understanding. We experiment with these models in five typical RL applications: ModelFail, ModelWin, and GridWorld, Fappybird, and SpaceInvaders-v0. We find that no single OPE method is consistently the best performer through our analysis, but hybrid-model-based methods are generally outperformed importance sampling methods. Most OPE estimators have strict constraints and do not perform well in Horizon length. Compared with DM estimators in directed-model-based methods, WDR, MAGIC, and RLTMLE estimators in hybrid-model-based methods are not always optimal in the ModelWin environment. We also notice that MAGIC and RLTMLE perform well in most cases. Still, when the historical trajectory data are relatively less, the MSE between the estimated state value of the target policy and the state value of the actual execution of the target policy is relatively big. Finally, we present future challenges and possible developing directions of OPE. Due to the difference between behavior policy and target policy and possible reward sparseness of behavior policy in some emerging applications, OPE still has a lot of challenges on its way. Happily noting that Google's PU-learning method breaks some OPE strict constraints, giving inspiration for future OPE researches and grounding OPE applications.

Keywords　artificial intelligence; reinforcement learning; off-policy evaluation; importance sampling; PU learning

1　引　言

强化学习(Reinforcement Learning,RL)作为一种自主学习技术,与监督学习、无监督学习共同支撑起整个机器学习框架.近年来,强化学习正在推动人工智能在机器人、自动驾驶等领域落地,如波士顿动力的机器人[1]和 AlphaGo Zero[2].强化学习的自主性依赖于智能体与环境交互所获的奖励反馈,通过不断试错的方式进行学习.

智能体与真实物理环境的交互有时是昂贵的和风险巨大的(如自动驾驶、医学治疗),而构建一个高精度的环境模拟器通常又极为困难.因此,学术界和工业界致力于离线策略评估(Off-Policy Evaluation,OPE)的研究,即研究如何对一个目标强化学习策略(简称目标策略)的评估只使用由不同的强化学习策略实际执行产生的历史数据,而并不实际运行目标策略.其中,上述实际执行策略也被称为行为策略.借助离线策略评估,研究人员可以在某个强化学习策略实际部署前进行评估,实现策略的提前过滤,从而避免强化学习策略的盲目部署.经过二十多年发展,离线策略评估正在产生巨大的应用价值,如医药、金融、广告和教育等[3-6].

离线策略评估最早的应用是上下文赌博机(Contextual Bandits,CB)[7].在该应用中,智能体反复观察上下文,选择动作,并获得奖励.在基于上下文赌博机的新闻推荐系统中,上下文包括用户访问历史信息,动作是文章推荐,奖励触发为用户点

击[8]. 在上述应用中,智能体的当前行为只会影响其到达下一个状态时所获得的奖励大小,而不会对未来状态的奖励值产生影响,属于单步决策过程.

离线策略评估主要应用于基于马尔可夫决策过程(Markov Decision Process,MDP)的强化学习中. 不同于基于上下文赌博机的单步决策过程,强化学习智能体的当前行为不仅会影响其到达下一个状态时所获得的奖励大小,还会对未来状态的奖励值产生影响,属于多步决策的过程. 由于多步决策,评估得到的状态价值方差会随着决策步数的增加而呈指数级增长,这成为离线策略评估在强化学习应用中面临的新挑战.

离线策略评估的目标是,使得目标策略状态价值的评估值与实际执行所获的真实值,均方误差最小. 实现该目标的前提是,行为策略产生的状态-动作组合应能覆盖目标策略所产生的状态-动作组合. 然而在实际应用中,用来评估的目标策略有可能会在某个状态选择一个从未在历史轨迹中出现的新动作. 因此,如何减少行为策略与目标策略差异性对评估造成的影响,是当前离线策略评估热点研究问题之一.

离线策略评估方法主要包括四类:纯模型法、重要性采样法、混合模型法和 PU 学习法(Positive Unlabeled,PU),如表 1 所示. 纯模型法采用拟合行为策略的动力学模型,通过估计器计算目标策略的平均回报来获得状态价值评估. 大多情况下,纯模型法评估得到的目标策略状态价值与实际执行得到目标策略状态价值具有较小方差. 然而在行为策略产生的历史轨迹样本较少的情况下,纯模型法会导致目标策略状态价值的评估存在较大偏差[9]. 重要性采样法(Importance Sampling,IS)借鉴了数学领域的重要性采样方法,旨在纠正行为策略产生的轨迹分布与目标策略之间的不匹配[10-11]. 因此,重要性采样法评估得到的目标策略状态价值是无偏和强一致的,即评估得到的期望等于真实的期望. 然而由于引入了重要性权重,当行为策略与目标策略差异较大时,该方法评估得到的目标策略状态价值会具有较大方差[12]. 混合模型法结合了纯模型法与重要性采样法的优点,以双鲁棒模型(Doubly Robust,DR)为代表,其特点是状态价值偏差比纯模型法小,方差比重要性采样法小[12-14]. PU 学习法[15]是 2019 年 Google 针对机器人抓取应用场景提出的最新方法,该方法不同之处在于 PU 学习法利用半监督学习,将评估问题转换为分类问题,提出了基于 0-1 奖励的离线策略评估方法.

表 1 离线策略评估方法分类

方法	核心做法	优缺点	具体模型
纯模型法	利用所有轨迹数据来构造行为策略近似模型	较小方差,但可能出现较大偏差	值迭代[16],基于最大似然估计[17]
重要性采样法	利用所有轨迹数据的重要性加权平均来计算目标策略状态价值	可能出现较大方差,理论证明是无偏差	IS[10], Step-IS[10], WIS[10], Step-WIS[10], RIS[18], MIS[19], INCRIS[20]
混合模型法	纯模型法与重要性采样法相结合	较小方差,较小偏差	DR[7], WDR[14], MAGIC[14], LTMLE[21], RLTMLE[21]
PU 学习法	把评估转化为一个二分类任务	无需了解行为策略,但只适用于 0-1 奖励	OPC[15], SoftOPC[15]

本文主要描述了离线策略评估相关背景知识,对比分析各类离线策略评估方法的机理与模型,通过实验开展了主流离线策略评估模型的程序复现与性能对比. 最后总结了离线策略评估的技术挑战,并展望了未来发展方向.

2 预备知识

本节主要介绍离线策略评预备知识,涉及 MDP 与强化学习.

2.1 马尔可夫决策过程

马尔可夫性是指系统下一个状态 s_{t+1} 仅与当前状态 s_t 有关,而与之前状态无关. MDP 是在系统状态具有马尔可夫性的环境中模拟智能体的随机策略和回报[12]. MDP 由六元组构成$\langle S,A,R,P,P_0,\gamma\rangle$:

S:状态空间,$s_i\in S$ 为状态空间中第 i 个状态;

A:动作空间,$a_i\in A$ 为动作空间中第 i 个动作;

R:回报函数,$r(s'|s,a)$为在状态 s 时选择动作 a,转移到状态 s'所得到的奖励回报;

P:状态转移概率,$P(s'|s,a)$为在状态 s 时选择动作 a,转移到状态 s'的概率;

P_0:初始状态分布,$P_0(s)$为初始时刻选择状态 s 的概率;

γ:折扣因子,用于调整当前状态对后续状态回报值的影响,$\gamma\in[0,1]$.

在 MDP 动态过程中,智能体根据初始状态分布

P_0进入初始状态s_0，通过实施动作a_0，以$P(s'|s_0,a_0)$概率进入新状态s'，并反馈一个奖励回报$r(s'|s_0,a_0)$. 智能体在状态s'进一步采取新的动作，与环境持续交互.

2.2　强化学习基础

强化学习寻找的最优策略是从状态空间S到动作空间A的一个映射，可以表示为$\pi: S\times A\rightarrow[0,1]$，$\pi(a|s)=P[A_t=a|S_t=s]$是在状态$s$时采取动作$a$的概率. 策略累积回报可以表示为$R=\sum_{t=0}^{T}\gamma^t r_t$，可通过状态动作价值函数和状态价值函数计算.

状态动作价值函数$Q(s,a)$是在状态s处实施动作a所获累积回报期望. 在策略π下，$Q^\pi(s_t,a_t)=\mathbb{E}[r_{t+1}+\gamma r_{t+2}+\gamma^2 r_{t+3}+\cdots|s_t,a_t]$，简化为$Q^\pi(s_t,a_t)=\mathbb{E}[\gamma^{t'-t}r_{t'+1}|s_t,a_t]$，$t'\in[t,L-1]$，其中$L$是轨迹长度. 当动作选择遵循策略$\pi$，且以概率$P(s_{t+1}|s_t,a_t)$进入下一个状态$s_{t+1}$时，动作状态价值函数表示为

$$Q^\pi(s_t,a_t)=\sum_{t=0}^{L-1}\sum_{t'=t}^{L-1}(\gamma^{t'-t}r_{t'+1})P(s_{t+1}|s_t,a_t) \quad (1)$$

状态价值函数$V(s,a)$是在状态s处可获得的累积回报期望，可以使用贝尔曼方程(Bellman Equation)递归计算. 在策略π下，$V^\pi(s_t)=\sum_{a_t}\pi(a_t|s_t)Q^\pi(s_t,a_t)$，简化为$V^\pi(s_t)=\mathbb{E}[r_t+\gamma V^\pi(s_{t+1})]$.

离线策略评估是根据行为策略π_b实际执行轨迹数据D，来估计目标策略π_e的状态价值函数. D是由m条轨迹组成：$D=\{H_i,\pi_b\}_{i=1}^m$，其中，H_i为第i条轨迹. 每一条轨迹H是由一系列状态、动作和奖励对组成，表示为$H=(s_0,a_0,r_0,\cdots,s_{L-1},a_{L-1},r_{L-1})$，轨迹$H$的累积回报是$g(H)=\sum_{t=0}^{L-1}\gamma^t r_t$. 当指定第$i$条轨迹时，状态、动作、奖励分别表示为$s_t^i$，$a_t^i$，$r_t^i$.

离线策略评估的目标是最小化目标策略π_e的估计状态价值$\hat{V}(D)$与真实状态价值V^{π_e}的均方误差(MSE)[22]：

$$MSE(\hat{V}(D),V^{\pi_e})=\mathbb{E}[(\hat{V}(D),V^{\pi_e})^2] \quad (2)$$

本文中使用到的常用符号如表 2 所示.

表 2　常用符号表示及说明

符号	表示	符号	表示	符号	表示
π_b	行为策略	π_e	目标策略	M	MDP 模型
D	轨迹数据	H_i	第 i 条轨迹	L	轨迹长度
s_t^i	第 i 条轨迹 t 时刻状态	a_t^i	第 i 条轨迹 t 时刻动作	r_t^i	第 i 条轨迹 t 时刻奖励
$Q^\pi(s_t,a_t)$	策略 π 第 t 步的动作状态价值函数	$V^\pi(s_t)$	策略 π 第 t 步的状态价值函数	$g(H_i)$	轨迹 H_i 的累积回报
$\rho_t(H_i)$	每步重要性采样权重	w_t	平均累积重要比	$w_{0:t-1}^i$	第 i 个轨迹累积重要比

3　离线策略评估模型分类

本章节将从各类离线策略评估方法的机理、模型细节差异等方面，对纯模型法、重要性采样法、混合模型法和 PU 学习法进行详细阐述.

3.1　纯模型法

纯模型法是最早提出的离线策略评估方法. 在 MDP 六元组已知的情况下，该方法利用全部轨迹数据来估计目标策略的状态动作价值函数$Q^\pi(s,a)$. 其中，基于值迭代估计器[16]的离线策略评估算法最具代表性. 在 MDP 六元组未知情况下，常常采用基于最大似然估计的估计器[17]来预测回报函数R和状态转移概率P，并构造 MDP 的近似模型$\widetilde{M}$.

3.1.1　值迭代估计器

值迭代估计器可以有效地计算任意有限系统的状态价值[16]，将初始状态的状态价值设置为 0：$\widetilde{V}^{\pi_e}(s_0)=0$，对于所有的$s\in S$，$a\in A$，第$t$步的动作状态价值和状态价值分别计算为

$$\widetilde{Q}^{\pi_e}(s_t,a_t)=r_t+\gamma\sum_{s_{t+1}\in S}P(s_{t+1}|s_t,a_t)\widetilde{V}^{\pi_e}(s_{t+1}) \quad (3)$$

$$\widetilde{V}^{\pi_e}(s_t)=\sum_{a_t}\pi_e(a_t|s_t)\widetilde{Q}^{\pi_e}(s_t,a_t) \quad (4)$$

值迭代估计器所估目标策略的状态价值计算公式如下：

$$V_{\mathrm{DM}}=\frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{L-1}\pi_e(a_t^i|s_t^i)\widetilde{Q}^{\pi_e}(s_t^i,a_t^i) \quad (5)$$

当 MDP 未知时，该方法无法直接使用.

3.1.2　基于最大似然估计的估计器

基于最大似然估计的估计器，也称为直接法(Direct Model，DM)[17]，该方法分为两步. 首先从历史轨迹中拟合 MDP 近似模型$\widetilde{M}$；然后根据式(5)，使用所估参数$\widetilde{P}$和$\tilde{r}$估计目标策略的状态价值.

从历史轨迹中拟合 MDP 近似模型$\widetilde{M}$可通过最大似然估计计算. 在状态s、动作a处的估计值仅取决于历史轨迹子集$D_s^a=\{H_i\in D|s_t^i=s\wedge a_t^i=a\}$，$t\in[0,L-1]$. 回报函数$r(s,a)$的最大似然估计是

在所有历史轨迹中，处于状态 s 选择动作 a 所得奖励 r 的平均值，公式如下：

$$\tilde{r}(s,a)=\frac{1}{|D_s^a|}\sum_{H_i\in D_s^a} r_t^i(s,a) \tag{6}$$

其中，$|D_s^a|$ 表示包含状态 s 选择动作 a 的历史轨迹数量. 在状态 s 选择动作 a 并转移到状态 s' 的最大似然状态转移概率为

$$\tilde{P}(s'|s,a)=\sum_{H_i\in D_s^a, s_{t+1}^i=s'}\frac{1}{|D_s^a|}=\frac{|D_{ss'}^a|}{|D_s^a|} \tag{7}$$

其中，$D_{ss'}^a=\{H_i\in D_s^a \mid s_{t+1}^i=s'\}$，$|D_{ss'}^a|$ 表示在状态 s 选择动作 a 转移到状态 s' 的轨迹数量.

在历史轨迹中每个状态动作对都频繁出现情况下，直接法具有可证明的低方差和可忽略的偏差[9]. 但是现实应用通常具有较大状态空间，导致大多数状态无法被访问，或者只被访问一次. 此外，直接法是在不了解目标策略的情况下估计 MDP 的近似模型 $\widetilde{M}$，而历史轨迹可能集中在与目标策略无关且不充分的领域. 上述情况中，直接法并不适用.

除了值迭代估计器和直接法估计器之外，函数逼近也常常被应用于拟合 MDP 近似模型 $\widetilde{M}$[16,23] 和拟合值函数[24-25] 中. 当 MDP 六元组参数或策略的值函数不能准确表示时，会对估计值引入偏差，进而会破坏纯模型法所估可信度[26-28].

3.2 重要性采样法

重要性采样法借鉴了数学领域的重要性采样方法，旨在纠正行为策略产生的轨迹分布与目标策略之间的不匹配. 目前具有代表性的重要性采样估计器有：重要性采样估计器(IS)、加权重要性采样估计器(WIS)、回归重要性采样估计器(RIS)、边际重要性采样估计器(MIS)和增量重要性采样估计器(INCRIS).

3.2.1 重要性采样数学原理

重要性采样是处理分布不匹配的经典技术. 它从与原分布 f 不同的一个分布 f' 中采样，估计原分布 f 中随机变量 x 的期望值，基于以下公式：

$$\begin{aligned}\mathbb{E}_f[x]&=\int_x xf(x)\mathrm{d}x=\int_x x\frac{f(x)}{f'(x)}f'(x)\mathrm{d}x\\&=\mathbb{E}_{f'}\left[x\frac{f(x)}{f'(x)}\right]\approx\frac{1}{m}\sum_{i=1}^m x_i\frac{f(x_i)}{f'(x_i)}\end{aligned} \tag{8}$$

其中 x 是根据 f' 采集的样本，每个样本根据其在两个分布下发生的可能性比值进行加权. 这种加权更重视在采样分布 f' 下很少发生但在原分布 f 下频繁发生的样本. 重要性采样是强一致和无偏的. 其中强一致性是指当采集样本趋于无穷时，它以概率 1 收敛到 $\mathbb{E}_f\{x\}$；无偏是指在任意数量的样本下它的期望值都是 $\mathbb{E}_f\{x\}$[29].

加权重要性采样考虑了样本的平均累积重要比，加权重要性采样公式如下：

$$\frac{1}{m}\frac{\sum_{i=1}^m x_i\frac{f(x_i)}{f'(x_i)}}{\sum_{i=1}^m\frac{f(x_i)}{f'(x_i)}} \tag{9}$$

这种估计在实践中比重要性采样更快和更稳定. 例如当发生罕见事件时，两个分布下该事件发生的可能性比值将较大，那么重要性采样方法得到的估计量也会发生较大变化. 而加权重要性采样的分母起到了稳定估计量的作用.

3.2.2 重要性采样估计器

重要性采样估计器(IS)[10] 根据重要性采样方法对行为策略的回报进行重新加权，公式与重要性采样方法类似，表示如下：

$$V_{\mathrm{IS}}=\frac{1}{m}\sum_{i=1}^m g(H_i)\prod_{t=0}^{L-1}\rho_t(H_i) \tag{10}$$

其中，$\rho_t(H_i)=\pi_e(a_t^i|s_t^i)/\pi_b(a_t^i|s_t^i)$ 为每步重要性采样权重.

对每个步骤增加重要性采样权重称为逐步重要性采样权重估计器(Step-IS)公式如下：

$$V_{\text{step-IS}}=\frac{1}{m}\sum_{i=1}^m\sum_{t=0}^{L-1}\gamma^t r_t\prod_{t'=0}^{t}\rho_{t'}(H_i) \tag{11}$$

该估计器沿轨迹加权每个回报，当行为策略和目标策略相同时，$\rho_t(H_i)=1$，则估计器实际计算的是每条轨迹的平均回报.

在行为策略 π_b 已知，且对于所有状态动作对，满足在 $\pi_b(a|s)=0$ 时，$\pi_e(a|s)=0$，IS 和 Step-IS 均是无偏估计. 但是在许多应用中，行为策略 π_b 是未知的，在这种情况下，由于预测行为策略 π_b 会引入偏差，导致 IS 和 Step-IS 不再是无偏估计，其估计的状态价值方差会随轨迹长度增加而呈指数增长.

3.2.3 加权重要性采样估计器

加权重要性采样估计器(WIS)[10] 根据加权重要性采样，考虑了平均累积重要比(the average cumulative important ratio)，公式与加权重要性采样方法类似，表示如下：

$$V_{\mathrm{WIS}}=\frac{1}{m}\sum_{i=1}^m g(H_i)\frac{\prod_{t=0}^{L-1}\rho_t(H_i)}{w_{L-1}} \tag{12}$$

其中，$w_t=\frac{1}{m}\sum_{i=1}^m\prod_{t'=0}^{t}\rho_{t'}(H_i)$ 为所有轨迹第 t 步的平

均累积重要比.

与重要性采样估计器一样，加权重要性采样估计器也有对每个步骤都执行重要性采样权重的方法，该方法为逐步加权重要性采样估计器(Step-WIS)，公式如下：

$$V_{\text{step-WIS}}=\frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{L-1}\gamma^t r_t\frac{\prod_{t'=0}^{t}\rho_{t'}(H_i)}{w_t} \tag{13}$$

在行为策略π_b已知，且对于所有的状态动作对，满足在$\pi_b(a|s)=0$时，$\pi_e(a|s)=0$，加权重要性采样估计器是强一致的，但是有偏差，随着历史轨迹数量的增多，估计偏差会越来越小. Step-WIS 被认为是 IS、Step-IS、WIS、Step-WIS 中最实用的估计器[10].

3.2.4　回归重要性采样估计器

回归重要性采样估计器(RIS)[18]根据轨迹数据中状态动作的频率修正行为策略，使之与轨迹数据相符合，从而降低所估计的目标策略状态价值与目标策略真实执行的状态价值之间的均方误差.

在离线策略评估实际运用中，虽然轨迹数据 D 是基于π_b采样得到的，但由于采样误差的存在使得轨迹数据并不能准确匹配 π_b. 例如，π_b在特定状态下，选择两个动作的概率相等，但是历史轨迹数据可能显示选择一个动作的概率要高于选择另一个动作的概率. IS 估计器就会对该动作施加不正确的权重. 基于此，回归重要性采样估计器将采样动作的概率替换为历史轨迹中该动作实际发生的概率，即用经验估计代替行为策略 π_b.

回归重要性采样估计器定义 n 步轨迹，$H_{t-n:t}$：$(s_{t-n},a_{t-n},\cdots,s_{t-1},a_{t-1},s_t)$，为轨迹 H 部分轨迹，当 $t-n<0$ 时，$H_{t-n:t}$表示轨迹的前 t 步. 策略 $\pi,\pi\in\Pi^n,\pi:S^{n+1}\times A^n\to[0,1]$，为 n 步轨迹 $H_{t-n:t}$对应的策略. 然后估计 Π^n的最大似然行为策略，公式如下：

$$\pi(D)^{(n)}=\arg\max_{\pi\in\Pi^n}\sum_{H\in D}\sum_{t=0}^{L-1}\log\pi(a|H_{t-n:t}) \tag{14}$$

与重要性采样估计器不同，该方法权重考虑状态 s_t、动作 a_t和部分轨迹 $H_{t-n:t}$，公式如下：

$$V_{\text{RIS}(n)}=\frac{1}{m}\sum_{i=1}^{m}g(H_i)\prod_{t=0}^{L-1}\frac{\pi_e(a_t^i|s_t^i)}{\pi(D)^{(n)}(a_t^i|H_{t-n:t}^i)} \tag{15}$$

在确定性环境中，采样误差仅来自采样动作，可以通过经验估计来纠正. 但在随机环境中，采样误差还包括状态和奖励. 纠正状态和奖励引起的采样误差需要知道真实状态和奖励频率，但在离线策略评估设置中，这些频率通常是未知的. 此外，该方法假设 D 至少包含一条在π_b下可能的轨迹，如果说有任何一条 π_b下可能的轨迹在 D 中不存在，那么理论证明 RIS($L-1$)就不再是无偏估计器.

3.2.5　边际重要性采样估计器

边际重要性采样估计器(MIS)[19]基于马尔可夫独立性假设，利用经验平均计算边际状态分布，使离线策略评估状态价值方差与轨迹长度不存在指数依赖.

在时间 t 处的边际状态分布表示如下：$d_t^{\pi_b}(s_t)$和 $d_t^{\pi_e}(s_t)$. 边际状态分布 $d_t^{\pi_b}(s_t)$和 $d_t^{\pi_e}(s_t)$是关于策略的函数，是递归定义的动态函数，对于 π_b(π_e类似)定义为

$$d_t^{\pi_b}(s_t)=\sum_{s_{t-1}}P_t^{\pi_b}(s_t|s_{t-1})d_{t-1}^{\pi_b}(s_{t-1}) \tag{16}$$

其中，$P_t^{\pi_b}(s_t|s_{t-1})$利用 MDP 的状态转移与策略的乘积计算，公式如下：

$$P_t^{\pi_b}(s_t|s_{t-1})=\sum_{a_t\in A}\pi_b(a_t|s_{t-1})P(s_t|s_{t-1},a_{t-1}) \tag{17}$$

当 $t=1$ 时，行为策略的边际状态分布和目标策略的边际状态分布相同：$d_1^{\pi_b}=d_1^{\pi_e}=d_1$. 根据边际状态分布的定义，离线策略评估估计目标策略的状态价值可表示为

$$V^{\pi_e,L}=\sum_{t=0}^{L-1}\sum_{s_t}d_t^{\pi_e}(s_t)\sum_{a_t}\pi_e(a_t|s_t)\sum_{s_{t+1}}P_{t+1}(s_{t+1}|s_t,a_t)r_t(s_t,a_t,s_{t+1}) \tag{18}$$

在离线策略评估环境中，一般只知道行为策略和目标策略，无法得知上述公式中的 $r(s_t,a_t,s_{t+1})$以及策略变化所隐含的状态分布 $d_t^{\pi_e}(s_t)$，$\forall t>1$. 为了实现边际重要性采样估计器，Xie 等人[19]增加状态空间较小的约束，使得部分状态可直接观测，边际重要性采样估计器可计算如下：

$$V_{\text{MIS}}=\frac{1}{m}\sum_{i=1}^{n}\sum_{t=0}^{L-1}\frac{\widetilde{d}_t^{\pi_e}(s_t^i)}{\widetilde{d}_t^{\pi_b}(s_t^i)}\widetilde{r}_t^{\pi_e}(s_t^i) \tag{19}$$

其中，$\widetilde{d}_t^{\pi_b}(s_t^i)$采用经验均值：$\widetilde{d}_t^{\pi_b}(s_t^i)=\frac{1}{m}\sum_i\mathbf{1}(s_t^i=s_t)$，而且，当所有轨迹中都不存在状态 s_t时，定义 $\widetilde{d}_t^{\pi_e}(s_t^i)/\widetilde{d}_t^{\pi_b}(s_t^i)=0$. $\widetilde{d}_t^{\pi_e}(s_t^i)$和 $\widetilde{r}_t^{\pi_e}(s_t^i)$分别用以下公式计算：

$$\widetilde{d}_t^{\pi_e}=\widetilde{P}_t^{\pi_e}\widetilde{d}_{t-1}^{\pi_e} \tag{20}$$

$$\widetilde{P}_t^{\pi_e}(s_t|s_{t-1})=\frac{1}{m_{s_{t-1}}}\sum_{i=1}^{m}\frac{\pi_e(a_{t-1}^i|s_{t-1})}{\pi_b(a_{t-1}^i|s_{t-1})}\times\mathbf{1}((s_{t-1}^i,s_t^i)=(s_{t-1},s_t)) \tag{21}$$

$$\widetilde{r}_t^{\pi_e}(s_t^i)=\frac{1}{m_{s_t}}\sum_{i=1}^{m}\frac{\pi_e(a_{t-1}^i|s_{t-1})}{\pi_b(a_{t-1}^i|s_{t-1})}r_t^i\mathbf{1}(s_t^i=s_t) \tag{22}$$

其中，$\widehat{P}_t^{\pi_e}$ 和 $\tilde{r}_t^{\pi_e}$ 将重要性采样权重与经验均值相结合，m_{s_t} 是在时刻 t，状态 s_t 的经验访问频率.

MIS 估计器适用于状态空间较小的环境，通过计算边际状态分布，来降低所估计的目标策略状态价值与目标策略真实执行的状态价值间的均方误差.

3.2.6 增量重要性采样估计器

增量重要性采样估计器(INCRIS)[20]特别针对基于选项(option)的策略[30]，在离线策略评估中引入时间抽象.

时间抽象可以降低规划和在线学习的计算复杂度，一种流行形式是使用子策略，特别是基于选项的策略. 选项可以表示为一个三元组$\langle I,\pi,\beta\rangle$，$\pi:S\times A\to[0,1]$表示一个选项的策略，$\beta:D\to[0,1]$表示终止条件，$\tau\in D$ 表示从该选项开始的部分轨迹，$\beta(\tau)$表示终止并退出该选项的概率. $I\subseteq S$ 表示选项的初始状态集合. 例如，一个名为“开门”的选项：π：包括一个到达、抓取和转动门旋钮的策略，β：一个识别门已经打开的终止条件，I：想要开门到有门存在的状态.

基于选项的策略将定义更高层次策略和更高层次的 k 步轨迹. 更高层次策略 μ[24]，$\mu(o_t\mid s_o,a_0,\cdots,s_t,a_t)$表示根据策略 μ，在给定历史轨迹$(s_o,a_0,\cdots,s_t,a_t)$时，选择选项 o_t的概率. 更高层次的 k 步轨迹表示为 $T=(s_o,o_0,v_0,\cdots,s_k,o_k,v_k)$，其中 v_t为选择选项 o_t得到的累计奖励. 对于选项缺失情况，将行为策略 μ_b或目标策略 μ_e中的缺失选项概率设置为零. 在该设置下，可以很容易地将 Step-IS 估计器应用于基于选项的策略，公式如下：

$$V_{\text{options-base}}=\frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{k^i}x_t^i y_t^i \tag{23}$$

与 Step-IS 估计器相比，x_t^i与 $\rho_t(H_i)$类似，只不过是更高层次策略的比值，y_t^i与奖励值类似，但分为基本策略相同和不同两种情况，当基本策略相同时，直接采用更高层次轨迹的累计奖励，当基本策略不同时，累计奖励需要考虑重要性权重，即基本策略的比值，公式如下：

$$x_t^i=\prod_{u=1}^{t}\frac{\mu_e(o_u^i\mid s_u^i)}{\mu_b(o_u^i\mid s_u^i)} \tag{24}$$

$$y_t^i=\begin{cases}v_t^i, & \text{当 } o_t^i\in O\\ \sum_{b=1}^{j_t^i}\rho_{t,b}^i r_{t,b}^i, & \text{当 } o_t^i\in\bar{O}\end{cases} \tag{25}$$

$$\rho_{t,b}^i=\prod_{c=1}^{j_t^b}\frac{\pi_e(a_{t,c}^i\mid s_{t,c}^i,o_t^i)}{\pi_b(a_{t,c}^i\mid s_{t,c}^i,o_t^i)}$$

其中 $r_{t,b}^i$是选项 o_t^i子轨迹中第 b 个回报，对于 s 和 a 也是如此；O 是一组选项，它们在 μ_b 和 μ_e之间具有相同的基本策略；$\bar{O}$ 是一组已更改基本策略的选项；k^i是数据集 D 的第 i 条轨迹的高层次轨迹的长度；j_t^i是选项 o_t^i产生的子轨迹的长度.

通常，选项用于实现特定的子任务，例如，在机器人导航任务中，有一个选项可以导航到一个固定位置. 随着不断的实验，会发现有一种更快的方法可以导航到该位置，所以人们会改变这个选项，并试图评估新的策略，验证其是否更好. 在这种情况下，旧选项和新选项都能够到达该固定位置，而唯一区别是新选项到达速度更快. 此时，选项被称为固定选项，无论行为策略还是目标策略，在选项终止的状态分布是相同的. 由于轨迹数据可被分为多个轨迹段，离线策略评估的状态价值可计算为每个轨迹段状态价值的总和. 例如，o_1 和 o_2 是固定选项，h_1是轨迹的第一段，直到并包括 o_1第一次发生，h_2 是第一次发生 o_1后直到并包括第一次发生 o_2 的轨迹部分，则：

$$\begin{aligned}\mathbb{E}_{\mu_e}(g(h))&=\mathbb{E}_{\mu_b}(V_{\text{step-IS}}(h))\\&=\mathbb{E}_{\mu_b}(V_{\text{step-IS}}(h_1))+\mathbb{E}_{\mu_b}(V_{\text{step-IS}}(h_2))\end{aligned} \tag{26}$$

固定选项可以看作是从重要性采样估计器中丢弃某些重要性采样权重的一种形式. 在估计固定选项之后的奖励时，可以将固定选项之前的权重丢弃. 该方法称之为增量重要性采样估计器，公式如下：

$$V_{\text{INCRIS}}=\frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{L-1}r_t\prod_{t'=t-k+1}^{t}\rho_{t'}(H_i) \tag{27}$$

其中，使用协方差检验优化 k，以降低 MSE.

INCRIS 估计器是强一致的，且均方误差在一些历史轨迹数量下，比 IS 估计器实现了两个数量级的改进. 但该估计器适用的环境必须有固定的子任务，且行为策略和目标策略在该子任务终止的状态分布相同.

对上述几种方法进行总结，通过统一的公式形式比对各个方法的异同点：

$$V(D)=\frac{1}{m}\sum_{i=1}^{m}R^i W^i \tag{28}$$

上述几种方法的优缺点对比如表 3 所示，对重要性采样估计器的优化都是通过改变重要性采样权重，目的是使方差得到降低，进而改变均方误差，优化离线策略评估.

表 3　重要性采样方法对比

方法	R^i	W^i	优缺点	验证场景
IS	$\sum_{t=0}^{L-1}\gamma^t r_t^i$	$\prod_{t=0}^{L-1}\rho_t(H_i)$	强一致，无偏差，但方差与轨迹长度呈指数级增长	历史轨迹长度相对较短
Step-IS	$\sum_{t=0}^{L-1}\gamma^t r_t^i$	$\prod_{t'=0}^{t}\rho_{t'}(H_i)$		
WIS	$\sum_{t=0}^{L-1}\gamma^t r_t^i$	$\prod_{t=0}^{L-1}\rho_t(H_i)/w_{L-1}$	强一致，有偏差，对于不太可能发生的事件，可以防止其权重太大	
Step-WIS	$\sum_{t=0}^{L-1}\gamma^t r_t^i$	$\prod_{t'=0}^{t}\rho_{t'}(H_i)/w_{t'}$		
RIS	$\sum_{t=0}^{L-1}\gamma^t r_t^i$	$\prod_{t=0}^{L-1}\frac{\pi_e(a_t^i\mid s_t^i)}{\pi(D)^{(n)}(a_t^i\mid H_{t-n:t}^i)}$	有偏差，降低了采样误差	历史轨迹相对覆盖面广
MIS	$r_t^{\pi_e}(s_t^i)$	$\sum_{t=0}^{L-1}\frac{\hat{d}_t^{\pi_e}(s_t^i)}{\hat{d}_t^{\pi_b}(s_t^i)}$	有偏差，方差小，方差与轨迹长度不呈指数依赖	整个环境的状态空间比较小
INCRIS	$\sum_{t=0}^{L-1}r_t$	$\prod_{t'=t-k+1}^{t}\rho_{t'}(H_i)$	强一致，均方误差降低，环境要求严格	历史轨迹有明显分段

3.3　混合模型法

混合模型法是将纯模型法和重要性采样法的优点相结合，形成一个方差小与偏差小的估计器. 目前有许多混合模型法的估计器，其中代表性的有：双鲁棒估计器(DR)、加权双鲁棒估计器(WDR)、模型和导向重要性采样联合估计器(MAGIC)、纵向目标最大似然估计器(LTMLE).

3.3.1　双鲁棒估计器

双鲁棒估计器(DR)[7]将纯模型法(DM)和重要性采样法中的逐步重要性采样估计器(Step-IS)相结合，克服单独使用 DM 估计器存在偏差大和单独使用 Step-IS 估计器存在方差大的问题.

双鲁棒估计[31-35]是一种从具有重要性质、不完全数据中进行估计的统计方法. DR 估计器引入双鲁棒估计增加获得可靠估计的机会，当两个估计器中任何一个是无偏的，DR 估计器即为无偏估计器.

双鲁棒估计器将 DM 估计器和 Step-IS 估计器结合在一起，公式如下：

$$V_{\mathrm{DR}}=\frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{L-1}\gamma^t w_{0:t}^i r_t^i-\frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{L-1}\gamma^t\left(w_{0:t}^i\widetilde{Q}^{\pi_e}(s_t^i,a_t^i)-w_{0:t-1}^i\widetilde{V}^{\pi_e}(s_t^i)\right)\tag{29}$$

其中，$w_{0:t-1}^i=\prod_{t'=0}^{t}\rho_{t'}(H_i)$ 是累积重要比(the cumulative importance ratio).

DR 估计器的偏差是 DM 估计器偏差和 Step-IS 估计器偏差的乘积，当 DM 估计器和 Step-IS 估计器至少存在一个是无偏时，DR 估计器则为无偏估计器[12]. 为验证以上结论，Jiang、Li[13]和 Thomas、Brunskill[14]在行为策略已知假设下，分别理论证明了在轨迹长度有限和轨迹长度无限的情况下，DR 估计器是无偏估计. 但当行为策略未知时，DR 估计器就不再是无偏估计器.

此外，DR 估计器根据实证结果表明可以在不引入偏差的情况下显著降低重要性采样估计器(IS)和逐步重要性采样估计器(Step-IS)的方差.

3.3.2　加权双鲁棒估计器

加权双鲁棒估计器(WDR)[14]将纯模型法(DM)和重要性采样法中的逐步加权重要性采样估计器(Step-WIS)相结合.

WDR 估计器基于 DM 估计器和 Step-WIS 估计器，公式如下：

$$V_{\mathrm{WDR}}=\frac{1}{m}\sum_{i=1}^{m}\widetilde{V}^{\pi_e}(s_0^i)+\sum_{i=1}^{m}\sum_{t=0}^{L-1}\gamma^t w_t^i\left[r_t^i-\widetilde{Q}^{\pi_e}(s_t^i,a_t^i)+\gamma\widetilde{V}^{\pi_e}(s_{t+1}^i)\right]\tag{30}$$

其中，$w_t^i=\prod_{t'=0}^{t}\rho_{t'}(H_i)\Big/\sum_{j=1}^{n}\prod_{t'=0}^{t}\rho_{t'}(H_j)$.

WDR 估计器通常优于其他重要性采样估计器，但在一些环境中，DM 估计器可能产生更低的 MSE 估计，这一结果在 Thomas、Brunskill[14]的实验中被发现.

3.3.3　模型和导向重要性采样联合估计器

模型和导向重要性采样联合估计器(MAGIC)[14]将纯模型法(DM)和加权双鲁棒估计器(WDR)相结合. MAGIC 估计器首先将每一条轨迹分为两个部分(第 0 到 j 步，第 j 步到 $L-1$ 步)，轨迹前一部分(第 0 到 j 步)采用 WDR 估计器，后一部分(第 j 步到 $L-1$ 步)采用 DM 估计器，通过计算不同长度

部分轨迹的回报加权平均值实现 MSE 最小化.

$g^{(j)}(D)$是参数为 j 的估计值,计算公式如下:

$$g^{(j)}(D)=\mathrm{WDR}^{[0:j]}(D)+\mathrm{DM}^{[j:L-1]}(D) \quad (31)$$

当 $j=0$ 时,该方法是一个纯粹的 DM 估计器,当 $j=L-1$ 时,该方法是一个纯粹的 WDR 估计器,而当 $0<j<L-1$ 时,该方法是混合 DM 和 WDR 的估计器.当 j 较小时,此时 MAGIC 估计器类似于 DM 估计器,方差小偏差大.当 j 较大时,此时 MAGIC 估计器类似于 WDR 估计器,方差大偏差小.因此,随着 j 的增加,方差逐渐增加,偏差逐渐减小.

当历史轨迹数据 D 中有 m 条轨迹时,参数为 j 的估计值用以下公式表示:

$$g^{(j)}(D)=\sum_{i=1}^{m} g_i^{(j)} \quad (32)$$

MAGIC 估计器的最终估计量是不同参数 j 的估计值 $g^{(j)}(D)$的凸组合.理想情况下,希望这个凸组合最小化均方误差(MSE),它通过计算不同长度回报的加权平均值来实现:$(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{g}$,其中 $\boldsymbol{g}=(g^{-1}(D),\cdots,g^{L-1}(D))$.

在轨迹长度、状态和动作有限,且对于所有的状态动作对,满足在 $\pi_b(a\mid s)=0$ 时,$\pi_e(a\mid s)=0$,MAGIC 估计器被理论证明是强一致的.但当这些条件不满足时,这并不意味着 MAGIC 估计器的性能很差,仅仅意味着理论结果没有得到保证.

3.3.4 纵向目标最大似然估计器

纵向目标最大似然估计器(LTMLE)[21]将历史轨迹分为两部分:$D^{(0)}=(H_1,\cdots,H_{(1-p)m})$和 $D^{(1)}=(H_{(1-p)m+1},\cdots,H_m)$,$0<p<1$.通过拟合 MDP 或 SARSA 等方法[22],用 $D^{(0)}$估计动作状态价值函数,形成初始估计.以这种方式得到的初始估计往往表现出较小的方差,但偏差较大.LTMLE 估计器利用纵向目标最大似然估计[36-37],在初始估计拟合的基础上定义了一个参数模型 $\hat{Q}^{\pi_e,\delta}(\varepsilon_t)$,$\forall\delta\in(0,1/2)$,称为第二阶段参数模型,并在样本 $D^{(1)}$ 上,通过最大似然拟合该参数模型,减小偏差.

对于所有的 $s_t\in S$,$\hat{V}^{\pi_e}(\varepsilon_t)(s_t)$作为与第二阶段参数模型对应的状态价值函数,计算公式如下:

$$\widehat{V}^{\pi_e}(\varepsilon_t)(s_t)=\sum_{a_t\in A}\pi_e(a_t\mid s_t)\widehat{Q}^{\pi_e,\delta}(\varepsilon_t) \quad (33)$$

LTMLE 估计器公式如下:

$$V_{\mathrm{LTMLE}}=\widehat{V}^{\pi_e}(\varepsilon_1)(s_1) \quad (34)$$

LTMLE 估计器假设所有轨迹具有相同的初始状态 s_1.为解决方差不稳定问题,LTMLE 又引入三种正则化技术:软化权重、部分轨迹长度、惩罚,提出了改进的 RLTMLE 估计器.RLTMLE 估计器允许使用重要性采样权重,而不是直接对其求和,不引入偏差的情况下提高了稳定性.

对上述几种方法进行总结,比较如表 4 所示.

表 4 混合模型方法对比

方法	原理	优缺点	验证场景
DR	DR 估计器将 DM 和 IS 结合在一起,DM 被用来指导减少方差,但不是完全取代重要性采样估计	如果 DM 和 IS 中的任何一个是无偏的,那么估计是无偏的.但是 DR 只减少了动作随机性的方差	历史轨迹长度相对较短
WDR	WDR 估计器是在 DR 估计器中应用一个简单的已知的重要性采样估计扩展 WIS	如果近似模型是完美的,WDR 是一个很好的估计器,但如果状态转移或奖励是随机的,则无法保证性能	
MAGIC	MAGIC 结合 WDR 和 DM,通过前几步采用 WDR 估计,后面的采用 DM 估计得到部分重要性采样估计,通过计算不同长度回报的加权平均值来最小化 MSE	强一致的,均方误差减小,但不适用于平均奖励	历史轨迹相对覆盖面广
LTMLE	LTMLE 根据纵向目标最大似然方法,估计状态价值函数	强一致的,提高了稳定性,但它要求历史轨迹初始状态相同	环境奖励有上下界
RLTMLE	在 LTMLE 估计器引入软化权重、部分轨迹长度、惩罚正则化		

3.4 PU 学习法

PU 学习法[15]是 2019 年 Google 最新提出的方法.与之前方法不同之处在于 PU 学习法利用半监督学习的思路,针对行为策略奖励稀疏性提出基于 0-1 奖励的离线策略评估方法.

PU 学习法是一种基于离线策略分类(Off-Policy Classification,OPC)的新型离线策略评估方法[15].该方法将评估当作一个分类任务,根据历史轨迹对状态和动作进行基于半监督学习的二分类(可能导致成功或一定导致失败),目标策略的好坏取决于导致成功的状态与动作比例.

OPC 主要建立在两个假设之上:(1)稳定的 MDP;(2)智能体在每一轮实验要么成功要么失败.这些任务非常常见,例如,拾取物体、走迷宫、赢得游戏等.由于要么成功要么失败,所以每个状态和动作具有二分类标签.如果某一个动作是可以导致成功,那么就将其称为有效的动作;而如果某一个动作一定会导致失败,那么就将其称为灾难性的动作.同

理，如果某一个状态所采取的动作有一个可以导致成功，那么就将其称为有效的状态；而如果某一个状态的所有动作都会导致失败，那么就将其称为灾难性的状态.

当某一轮实验成功时，该轮实验中所有状态与动作是有效的. 而当某一轮失败时，无法知道具体哪个动作或状态导致失败. 因此，OPC 可利用 PU 学习[38]，在只有正类和无标记数据情况下，训练二分类器，进行有效策略评估，公式如下：

$$\mathrm{OPC}(Q)=p(y=1)\mathbb{E}_{(s,a),y=1}[1_{Q(s,a)>b}]-\mathbb{E}_{(s,a)}[1_{Q(s,a)>b}] \tag{35}$$

其中，y 是类别，$y=1$ 是有效动作，Q 是状态动作价值函数. 为了对每个 Q 函数公平，每个 Q 函数单独设置门槛 b，以最大化 OPC(Q).

此外，还可以采用另一种方法 SoftOPC，公式如下：

$$\mathrm{SoftOPC}(Q)=p(y=1)\mathbb{E}_{(s,a),y=1}[Q(s,a)]-\mathbb{E}_{(s,a)}[Q(s,a)] \tag{36}$$

PU 学习法不需要了解行为策略，适用于更多环境. 此外，PU 学习法不需要重新调整轨迹数据权重，因此方差不会随着轨迹长度的增加而变大. 除此之外，PU 学习法还可以扩展到更大的任务，包括在现实世界中基于视觉的机器人抓取的任务. 在实验时，Irpan 等人[15]分别在机器人抓取任务的模拟环境和真实世界任务中测试 PU 学习法. 通过确定性系数 R^2[39] 和相关系统 ξ[40] 来评判 PU 学习法得分.

3.5　方法对比

针对纯模型法、重要性采样法、混合模型法和 PU 学习法的对比如表 5 所示.

表 5　离线策略评估方法对比

类别	方法	原理	优缺点
纯模型法	值迭代	利用贝尔曼方程	方法简单，但只适用于 MDP 已知
	DM	利用历史轨迹拟合 MDP 的近似模型	方差小，但有偏差
重要性采样法	IS、Step-IS	利用重要性采样原理	偏差小，但方差与轨迹长度呈指数级增长
	WIS、Step-WIS	利用加权重要性采样原理	方差小，不受小概率事件干扰，但有偏差
	RIS	利用轨迹数据中状态动作的频率修正行为策略	方差小，但有偏差
	MIS	利用经验均值，计算边际状态分布	方差小，不指数依赖轨迹长度，但有偏差
	INCRIS	利用高层次的轨迹及其策略	方差小，偏差小，但环境要求严格
混合模型法	DR	利用双鲁棒估计，结合 DM 和 IS	方差小，偏差小，但方差与轨迹长度呈指数级增长
	WDR	结合 DM 和 WIS	方差小，偏差小，不受小概率事件干扰
	MAGIC	结合 DM 和 WDR	方差小，偏差小，但只适用于非平均奖励的环境
	LTMLE、RLTMLE	利用纵向目标最大似然方法	方差小，偏差小，但历史轨迹初始状态必须相同
PU 学习法	OPC、SoftOPC	把评估当作一个二分类的任务	无需了解行为策略，但只适用于 0-1 奖励

纯模型法包括值迭代和基于最大似然估计. 值迭代可用于 MDP 已知时，计算方法简单. 而基于最大似然估计可用于 MDP 未知时，基于每个状态动作对频繁出现的历史轨迹拟合 MDP 近似模型，具有方差小的特点.

重要性采样法包括 IS、Step-IS、WIS、Step-WIS、RIS、MIS、INCRIS. 当历史轨迹长度较短时，IS 和 Step-IS 利用重要性采样原理，实现无偏估计. 当历史轨迹数量较多时，WIS 和 Step-WIS 利用加权重要性采样，实现强一致，但存在偏差. 当历史轨迹覆盖面较广时，RIS 根据轨迹数据中状态动作的频率修正行为策略，可降低采样误差. 当历史轨迹的状态空间较小时，MIS 利用相同状态转移模式的经验平均值，使所估计状态价值方差不指数依赖轨迹长度. 当历史轨迹具有明显分段时，INCRIS 可利用时间抽象，对基于选项的行为策略和目标策略，实现更准确的估计.

混合模型法包括 DR、WDR、MAGIC、LTMLE、RLTMLE. 当历史轨迹长度较短时，DR 结合 DM 和 IS 减少动作随机性导致的估计方差，WDR 结合 DM 和 WIS，防止小概率事件权重太大. 当历史轨迹覆盖面较广时，MAGIC 结合 DM 和 WDR，在非平均奖励的环境实现强一致估计. 当历史轨迹中奖励有上下界时，LTMLE 和 RLTMLE 利用纵向目标最大似然估计，具有较低的均方误差.

PU 学习法包括 OPC 和 SoftOPC，针对行为策略奖励稀疏性提出基于二分类任务的离线策略评估方法，该方法无需了解行为策略，但只适用于 0-1 奖励.

4　离线策略评估的应用

离线策略评估有很多应用，比如机器人、游戏、医药治疗等. 在这些应用中，使用离线策略评估一般

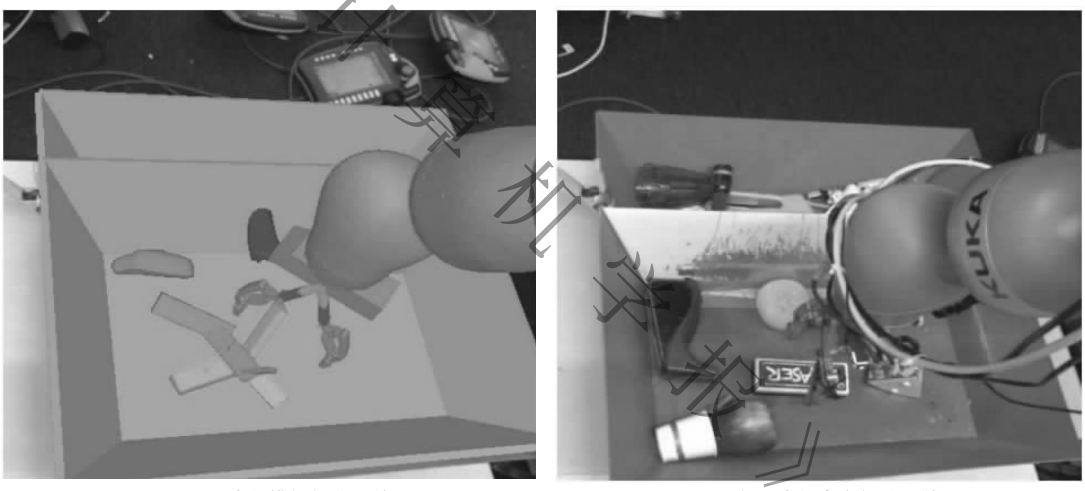
有两种方法.第一种方法是将离线策略评估应用在另一个领域收集的验证数据上,来衡量对新领域设置的泛化能力.第二种方法是在使用离线策略数据进行训练时,使用离线策略评估作为早期停止或模型选择的标准.

4.1 泛化能力

对于泛化能力,已经有论文[41-45]研究了深度强化学习的过拟合和记忆,提出了在模拟中定义测试环境来评估强化学习泛化能力.通常,很容易评估模拟环境中的策略,但对于真实世界的设置来说,这已经不够了,因为在真实世界中,模拟环境可能会非常昂贵,比如构建一个能够覆盖所有的医疗状况的模拟病人.因此,在现实问题中,离线策略评估是一个不可避免的部分,以一种有效、易于处理的方式度量泛化性能.

在强化学习中所面临的一些常见的泛化失败场景,比如离线策略训练数据不足且没有在线收集新数据时;训练数据是以系统偏差的方式生成的,而错过了部分目标分布时;训练和测试领域的差距等,原则上,只要根据从最终测试环境中采样的数据进行验证,所有这些场景都可以通过离线策略评估加以识别.

在机器人应用中,经常使用模拟环境来降低学习机器人技能的样本复杂度,由于模拟环境和真实环境的差距,策略的评估仍然需要使用一个真实机器人,但是使用真实机器人进行真值评估的效率非常低,因此使用离线策略评估提高评估的效率.PU学习法[15]针对上述泛化失败场景验证离线策略评估性能,特别是在机器人抓取任务中,使用真实世界(图1(b))数据评判模拟环境(图1(a))下训练的策略.实验结果表明,离线策略评估与多个环境的性能有很好的相关性.



(a)无风险的模拟抓取环境 (b)有风险的真实抓取环境

图1 离线策略评估在机器人抓取中的应用

在医药治疗应用中,目前已有很多研究[46-47]利用强化学习模拟患者与临床医生的动态交互.但由于模拟环境和真实世界间存在差距,将模拟环境中的策略应用于真实世界仍然存在较大风险,可使用离线策略评估降低风险.Li等人[48]在模拟环境中使用离线策略评估辅助重症监护病房败血症的药物治疗.

即使应用在相同场景中,不同离线策略评估方法的采样不尽相同.离线策略评估方法采样要求对比如表6所示.DM要求每个状态动作对频繁出现在历史轨迹中,因此需要较多历史轨迹.IS、Step-IS、WIS和Step-WIS的重要性采样权重的形式为累积和,估计的方差大小依赖于轨迹的长度,因此只对历史轨迹长度有要求.DR、WDR结合DM和重要性采样,在较多历史轨迹或历史轨迹长度较短时,估计较准确.MAGIC在非平均奖励环境下,需要较少历史轨迹.RLTMLE适用于初始状态确定的环境,需要较少历史轨迹.

表6 离线策略评估方法采样要求对比

方法	采样要求
DM	每个状态对频繁出现在历史轨迹中,需要较多历史轨迹
IS、Step-IS、WIS、Step-WIS	估计的方差大小依赖于轨迹的长度,需要较短历史轨迹,对轨迹数量无明显要求
DR、WDR	需要较短、较多历史轨迹
MAGIC	非平均奖励环境下,需要较少历史轨迹
RLTMLE	初始状态确定环境下,需要较少历史轨迹

4.2　模型选择标准

对强化学习方法的评价最直接的方法是通过与真实环境交互来评估已经学过的策略，但是，对于大多数现实场景，直接在真实环境中运行新策略的成本昂贵、风险巨大，甚至涉及到道德问题. 因此，使用离线策略评估作为早期停止或模型选择的标准是一个不可避免的部分.

在电子商务、搜索或新闻的推荐系统中，使用操作期间收集的数据测试新的内容服务算法[49]. 推荐系统评估的目的就是为了从多维度来评估该系统的实际效果和表现，从中发现可能的优缺点，通过优化该系统，为用户提供更优质的服务，获取更多的商业利益. 进行强化学习方法的评价最直接的方式是通过在线 A/B 测试来评估已经学到的策略，但是这种测试价格昂贵，而且可能会损害用户体验，因此，离线策略评估被用来评估不同推荐智能体的表现，如逆倾向分数(IPS)[50]、PI(Pseudoinverse estimator)[49]. 除此之外，Zou 等人[51]为提高用户的长期参与度，将信息流推荐形式化为一个马尔可夫决策过程，通过设计的 Q-网络来直接优化用户参与度，并通过离线策略评估比较不同算法的性能，进而选择最优策略.

在临床医学应用中，护理提供者筛选大量分散的多模式数据，进行临床决策评估，以降低患者风险、管理不确定性和控制成本. 当检查危重患者的临床决策时，如重症监护病房，这些压力会增加，需要快速判断，及时和适当的干预对于确保这些病例中患者的安全是至关重要的. Prasad[52]通过对重症护理干预的一系列案例研究，为临床医生循环决策支持提供了一个完整的框架，并通过离线策略评估进行奖励函数的设计. 临床部署前需要对学到的策略进行前瞻性的临床试验. 针对一种病症可能有多种方案，比如低血压的治疗并没有标准化，使用哪种血管加压素以及采用多少剂量并不清楚，而且对最佳方案的临床意见可能有显著不同. 如果直接在真实世界中进行评估，那么有可能造成无法估量的后果. 因此在实际部署前，应采用离线策略评估进行模型的选择. Futoma 等人[53]针对危重病人低血压的紧急情况，根据电子健康记录的观察数据提供多个策略，采用离线策略评估确定可行策略. Komorowski 等人[46]开发的 AI 临床医生，通过分析医生所做的上万次真实治疗决策，学习最佳的败血症治疗方案，AI 临床医生的数据流如图 2 所示，其中可以利用离线策略评估进行选择最优模型.

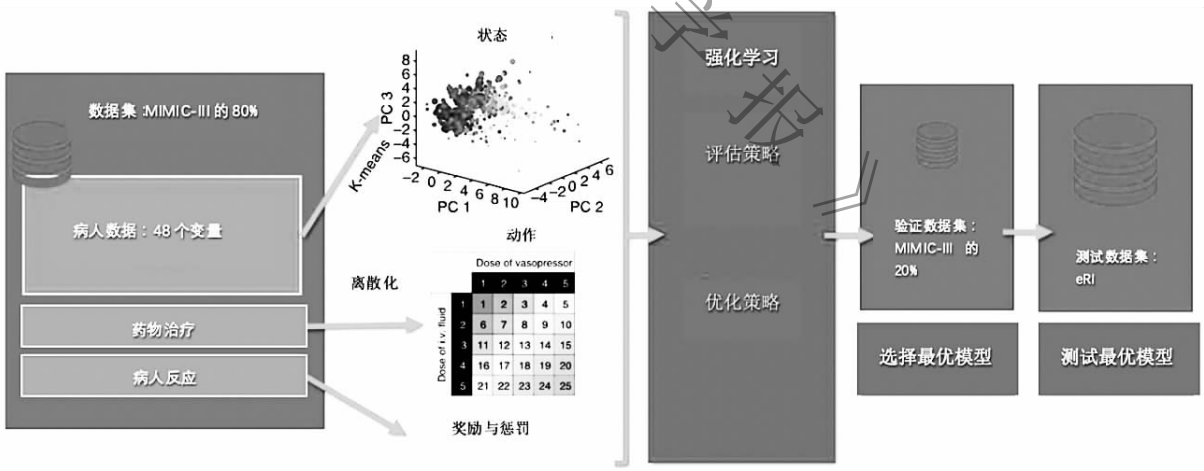


图 2　AI 临床医生的数据流

在多智能体应用中，如图 3 所示的 OpenAI 发布的 Neural MMO：大型多智能体游戏环境. Li 等人[54]利用离线策略评估方法提出一种离线强化学习算法，用于解决最佳同步问题. 将规定的行为策略应用于每个智能体，以生成和收集用于学习的历史轨迹数据. 为每个智能体导出一个离线策略的贝尔曼方程，以了解目标策略的价值函数，并同时找到改进的策略.

除此之外，离线策略评估还推动了完全的离策略强化学习的发展，使智能体完全使用历史数据进行学习. 使用之前智能体收集到的历史数据训练多个模型，利用离线策略评估从中选取最佳模型，进而在真实世界中测试该最优模型. 离线策略评估的使用，使评估效率增加，风险降低.

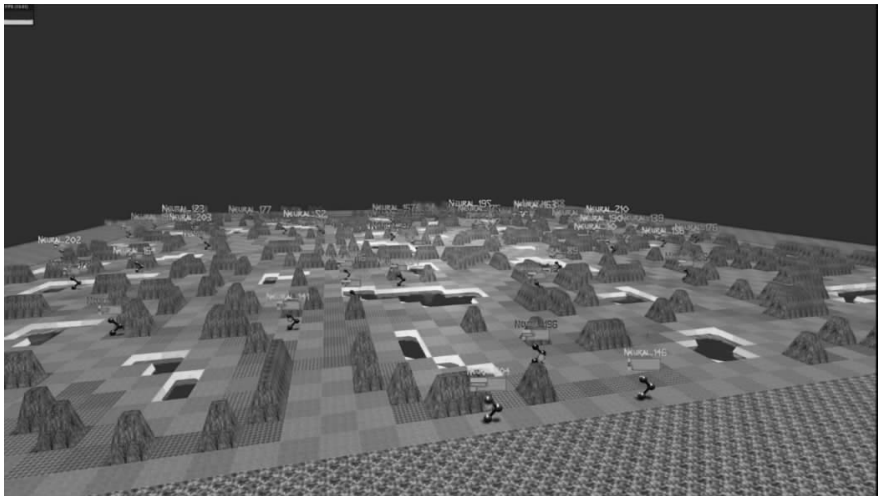


图 3 Neural MMO:大型多智能体游戏环境

5 实 验

本节通过实验尽最大可能复现了主流离线策略评估模型,并在 ModelFail、ModelWin、GridWorld 等领域论文常见的三个应用中进行了性能对比.实验的 OPE 涉及:直接法(DM)、重要性采样(IS)、逐步重要性采样(Step-IS)、加权重要性采样(WIS)、逐步加权重要性采样(Step-WIS)、双鲁棒(DR)、加权双鲁棒(WDR)、模型和导向重要性采样联合(MAGIC)、纵向目标最大似然(RLTLME).

5.1 物理环境

本节实验的软硬件配置如表 7 所示.

表 7 实验环境配置

操作系统	Windows10,64 位
CPU 型号	英特尔酷睿 i5-1035G1 双核
内存容量	16GB
硬盘容量	512GB
主要工具	R3.5.1; Rstudio1.2

5.2 实验环境

(1) ModelFail

如图 4 所示,ModelFail 的 MDP 有三个状态(用圆圈表示)和终端吸收状态(用双圈表示).智能体从状态 s_1 开始,它有两个动作,第一个动作 a_1 以概率 1 到达上层状态 s_2,奖励为 0.第二个动作 a_2 以概率 1 到达下层状态 s_3,奖励为 0.当智能体位于上层状态 s_2 时,采取动作 a_3 以概率 1 到达终端吸收状态 s_4,奖励为 1.当智能体位于下层状态 s_3 时,采取动作 a_4 以概率 1 到达终端吸收状态 s_4,但是得到负奖励,$r=-1$.

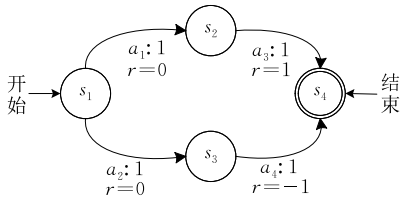


图 4 ModelFail MDP

设定实验中行为策略 a_1 的概率为 0.88,a_2 的概率为 0.12.目标策略则相反,它选择 a_1 概率为 0.12 和 a_2 的概率为 0.88.在轨迹采样时,轨迹长度为 10,轨迹数量 m 分别为 100、200、500、1000.

(2) ModelWin

如图 5 所示,ModelWin 的 MDP 有三种状态和一个终端吸收状态.智能体从状态 s_1 开始,它有两个动作 a_1 和 a_2.第一个动作 a_1 使智能体以概率 0.4 到达状态 s_2,奖励为 1,以概率 0.6 到达状态 s_3,奖励为−1.第二个动作 a_2 使智能体以概率为 0.6 到达状态 s_2,奖励为 1,以概率为 0.4 到达状态 s_3,奖励为−1.当智能体位于状态 s_2 和状态 s_3 时,以概率 1 到达终端吸收状态 s_1,奖励为 0.

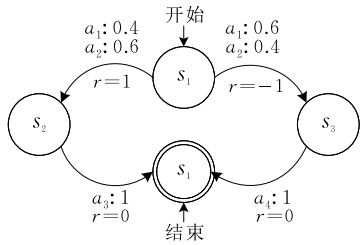


图 5 ModelWin MDP

实验时,行为策略在状态 s_1 采取动作 a_1 的概率为 0.73,动作 a_2 概率为 0.27,而目标策略则相反,它采取的动作 a_1 的概率约为 0.27,动作 a_2 的概率约为 0.73.在轨迹采样时,轨迹长度为 10,轨迹数量 m 分

别为 100，200，500，1000.

(3) GridWorld

GridWorld 代表了许多有共同属性的现实世界部分可观察的马尔可夫决策过程(Partially Observable Markov Decision Process，POMDP)[36]. 如图 6 所示，实验选用了 4×4 的 GridWorld，有 4 个动作，分别为：上、下、左、右. 在状态 s_6，智能体得到负奖励，$r=-10$，在状态 s_{14}，智能体得到正奖励，$r=1$，在状态 s_{16}，智能体得到正奖励，$r=10$，在其他状态，智能体得到负奖励，$r=-1$.

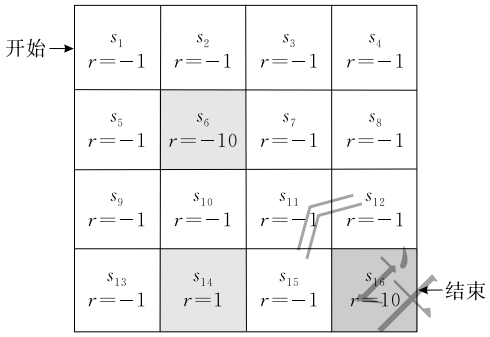


图 6　网格世界的图形描绘

实验中的行为策略和目标策略延用 Thomas 等人[55]提出的策略 π_4 和策略 π_5. 策略 π_4 是接近最优的策略，它通常采取动作向下移动，偶尔采取动作向右移动，迅速移动到 s_{14}，而且没有访问 s_6 得到 −10 奖励. 策略 π_5 是一个比策略 π_4 更优版本，减少动作向右移动的随机性. 设置轨迹长度 $T=100$，若在 $T=100$ 之前到达状态 s_{16}，则这轮结束，否则，直到 $T=100$ 结束，轨迹数量 m 分别为 100、200、500、1000.

(4) FlappyBird

FlappyBird 游戏，属于 PyGame Learning Environment，控制一只小鸟，跨越由各种不同长度水管所组成的障碍，如图 7 所示. 其状态空间三维，小鸟和下一个管道的水平距离、小鸟和底下管道的垂直距离、小鸟的速度，动作为两维一个人是小鸟飞行，一个是不做动作，小鸟会由于重力的作用而下降，奖励为小鸟过一个桩为 1，否则为 0. 设置轨迹长度 $T=500$，轨迹数量 m 分别为 100、200、500、1000.

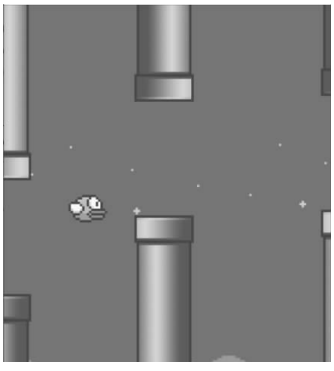


图 7　FlappyBird 游戏

(5) SpaceInvaders-v0

SpaceInvaders-v0 属于 openAI gym atari 游戏，如图 8 所示，太空入侵者中，支持的输入有 6 种，一个是什么也不做，一个是开火，另 4 个是控制方向：0：NOOP、1：FIRE、2：UP、3：RIGHT、4：LEFT、5：DOWN，如果你打掉 X 分的怪兽，奖励就为 X，状态是 RGB 图像(210，160，3)，动作空间 9 维. 每一个 episode，有三条命，如果三条命都被怪兽打死了，这个游戏就结束了. 设置轨迹长度 $T=1000$，轨迹数量 m 分别为 100、200、500、1000.

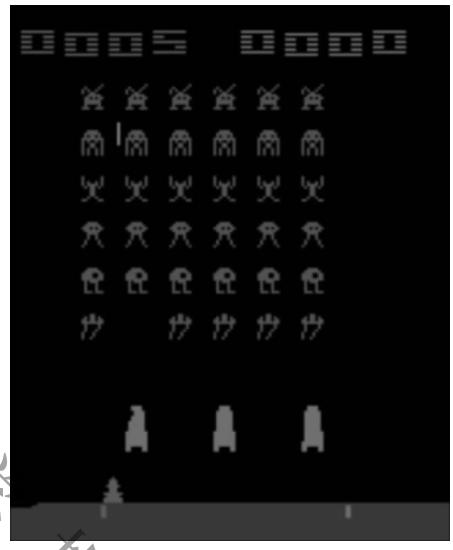


图 8　SpaceInvaders-v0 游戏

5.3　实验验证

OPE 方法的评估主要采用所估计的目标策略状态价值与目标策略真实执行的状态价值均方误差(MSE).

(1) ModelFail 结果

图 9 描述了在 ModelFail 上的结果，表 8 显示各估计器均方误差的具体值. 加权重要性采样方法，

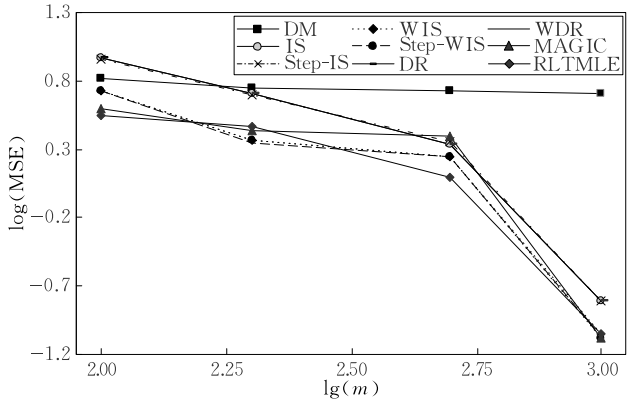


图 9　不同轨迹数量下的 ModelFail 实验结果

表 8 ModelFail 环境下估计器的性能对比

轨迹数量	DM	IS	Step-IS	WIS	Step-WIS	DR	WDR	MAGIC	RLTMLE
100	6.5890	9.20649	9.1564	5.3700	5.3514	9.20649	5.3215	3.93873	**3.5642**
200	5.5724	5.11548	4.9510	2.3500	2.3040	5.11548	**2.2258**	2.73838	2.9600
500	5.2890	2.17076	2.2560	1.7723	1.7632	2.17076	1.7562	2.48322	**1.2490**
1000	5.1563	0.15707	0.1543	0.0902	**0.0830**	0.15707	0.0852	0.08275	0.0890

WIS 和 Step-WIS 与 WDR 估计器的性能接近. 未加权的重要性采样方法 IS 和 Step-IS 与 DR 估计器性能接近.

随着历史轨迹数量的增加,每个估计器的 MSE 越来越低,其中 DM 估计器缓慢降低,逐渐趋于平稳. 当历史轨迹数量为 100 时,各估计器 MSE 都相对较高,最优估计器是 RLTMLE,MSE 为 3.5642,比最差的 IS 估计器和 DR 估计器实现了 61.3%的性能提升. 随着轨迹数量的增多,各个估计器性能逐渐提升,当历史轨迹数量为 1000 时,最优估计器为 Step-WIS,MSE 为 0.083,与 WIS 估计器、WDR 估计器、MAGIC 估计器和 RLTMLE 估计器相差不大,比最差的 DM 估计器实现了 98.3%的性能提升.

(2) ModelWin 结果

图 10 描述了在 ModelWin 上的结果,表 9 显示各估计器均方误差的具体值. 正如上面所分析的一样,DM 在近似模型快速收敛到真实的环境下,表现比较好.

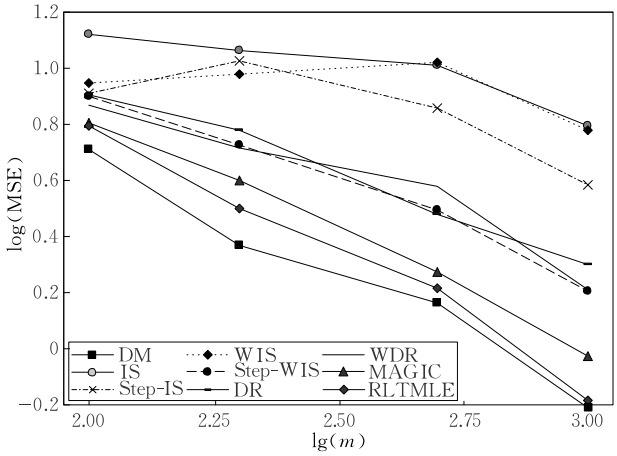


图 10 不同轨迹数量下的 ModelWin 实验结果

表 9 ModelWin 环境下估计器的性能对比

轨迹数量	DM	IS	Step-IS	WIS	Step-WIS	DR	WDR	MAGIC	RLTMLE
100	**5.12358**	13.25888	8.1523	8.83438	7.98837	8.00436	7.41197	6.39407	6.23319
200	**2.32450**	11.61523	10.5641	9.51298	5.33090	6.00501	5.18637	3.95971	3.15333
500	**1.45417**	10.25032	7.2364	10.54698	3.11286	3.00044	3.77160	1.87539	1.65140
1000	**0.61417**	6.26181	3.8452	5.97628	1.60696	2.00010	1.62732	0.93667	0.65210

不同的历史轨迹数量,DM 估计器都是其中最优的. 当历史轨迹数量为 100 时,DM 估计器的 MSE 为 5.12358,比最差的 IS 估计器实现了 61.3%的性能提升. 当轨迹数量为 1000 时,DM 估计器与 RLTMLE 估计器相差不大,比最差的 IS 估计器实现了 90.1%的性能提升.

(3) GridWorld 结果

图 11 描述了在 GridWorld 上结果,表 10 显示各估计器均方误差的具体值. DM 估计器在不同轨迹数量中表现最差,当历史轨迹数量为 1000 时,DM 估计器的 MSE 为 30.08946,数值非常高. 其中 WDR 估计器、MAGIC 估计器和 RLTMLE 估计器性能相近. 当历史轨迹数量为 100 时,最优估计器是 RLTMLE,MSE 为 3,比最差的 DM 估计器实现了 91.5%的性能提升. 当历史轨迹数量为 1000 时,最优估计器为 RLTMLE,MSE 为 1.20013,比最差的 DM 估计器实现了 96%的性能提升.

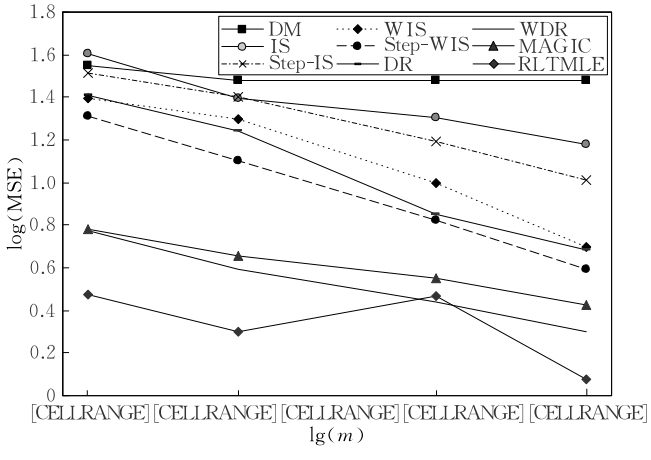


图 11 不同轨迹数量下的 GridWorld 实验结果

表 10　GridWorld 环境下估计器的性能对比

轨迹数量	DM	IS	Step-IS	WIS	Step-WIS	DR	WDR	MAGIC	RLTMLE
100	35.48946	40.01159	32.54450	25.00623	20.50450	25.51008	5.95265	6.00684	**3.00000**
200	30.09460	25.00242	25.34560	20.00049	12.71640	17.56047	3.91670	4.50836	**2.00238**
500	30.08946	20.01115	15.56015	10.00575	6.67456	7.10004	**2.75956**	3.55606	2.91844
1000	30.08946	15.02776	10.28238	5.00744	3.93286	4.84472	2.01005	2.66026	**1.20013**

(4) FlappyBird 结果

图 12 描述了在 FlappyBird 上的结果，表 11 显示各估计器均方误差的具体值. 重要性采样方法中的 IS、WIS 优于其他估计器. DM 估计器在不同轨迹数量中表现最差，当历史轨迹数量为 1000 时，DM 估计器的 MSE 为 19.49226，最优估计器 WIS 为 0.002258，两个估计器均方误差相差较大.

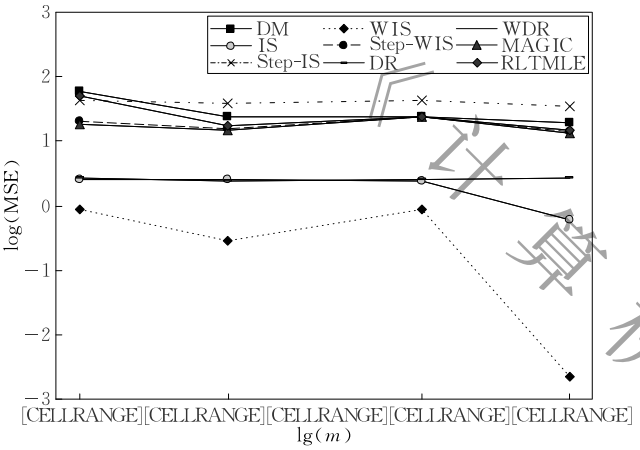


图 12　不同轨迹数量下的 FlappyBird 实验结果

(5) SpaceInvaders-v0 结果

图 13 描述了在 SpaceInvaders-v0 上的结果，表 12 显示各估计器均方误差的具体值. DM 估计器在不同轨迹数量中表现最差，当历史轨迹数量为 1000 时，DM 估计器的 MSE 为 28.08946，数值非常高. 重要性采样方法中的估计器与混合模型法中的估计器性能接近. 当历史轨迹数量为 100 时，几个估计器

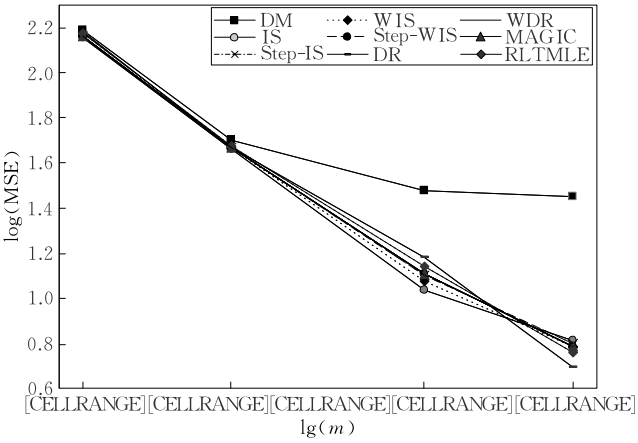


图 13　不同轨迹数量下的 SpaceInvaders-v0 实验结果

性能相差不大，MSE 都较高. 当历史轨迹数量为 1000 时，估计器的 MSE 都下降.

可以从这些实验中得出的初步结论是，WDR、MAGIC、RLTMLE 往往优于其他重要性采样估计器：IS、Step-IS、WIS 和 Step-WIS，以及混合模型法中的 DR. 但与 DM 估计器相比，WDR、MAGIC、RLTMLE 估计器并不总是最优的，如在 ModelWin 环境中，DM 估计器的性能优于所有估计器. 我们也注意到 MAGIC 和 RLTMLE 在历史轨迹数据比较少的时候，如 $n=100$ 时，所估计的目标策略状态价值与目标策略真实执行的状态价值均方误差比较大，导致这个结果的原因可能是因为这两种方法都需要协方差矩阵和偏差，但是当数据很少时，很难估计协方差矩阵，而且估计偏差时使用的一致性区间也是松散的.

表 11　FlappyBird 环境下估计器的性能对比

轨迹数量	DM	IS	Step-IS	WIS	Step-WIS	DR	WDR	MAGIC	RLTMLE
100	59.72764	2.604887	43.14255	**0.886873**	20.70969	2.64941	20.70969	18.72764	49.7719
200	24.40214	2.534901	38.05032	**0.291585**	15.90808	2.41282	15.90808	14.40214	17.1915
500	23.89722	2.443722	43.26375	**0.868682**	23.89722	2.54354	23.89722	23.89722	23.7719
1000	19.49226	0.595609	33.80989	**0.002258**	13.73256	2.70336	13.73256	13.49226	14.7719

表 12　SpaceInvaders-v0 环境下估计器的性能对比

轨迹数量	DM	IS	Step-IS	WIS	Step-WIS	DR	WDR	MAGIC	RLTMLE
100	154.89460	**143.350200**	145.419200	145.317000	145.879400	152.224400	145.879400	145.879400	149.7719
200	50.09464	**45.629110**	46.209710	46.245390	46.191410	46.180830	46.191410	46.191410	47.1915
500	30.08946	**10.830870**	12.642540	11.918990	12.841190	15.093200	12.841190	12.841190	13.7719
1000	28.08946	6.516912	6.263283	6.052279	6.125263	4.915947	6.125263	6.125263	**5.7719**

5.4 应用场景讨论

离线策略评估的应用是比较广泛的，涉及搜索引擎、推荐系统、广告投放系统、机器人、游戏、医药治疗等. 但是在公开发表的论文中，大多数对比实验选用 ModelWin、ModelFail、GridWorld 等小规模游戏场景. 谷歌将 PU 学习法应用于机器人抓取实验，其论文数据显示了离线策略评估方法在较大规模状态空间中的成功应用.

总体来说，离线策略评估的基线实验在大规模状态空间的应用环境并不是太多，这对该领域的研究是一个阻碍，亟待突破.

6 未来研究方向

尽管离线策略评估的研究已经引起了越来越多的关注，而且取得了一些成果，提出了很多相关理论和算法. 但从目前的研究和实用性来看，仍然存在一些挑战，可能形成未来主要的研究趋势.

(1) 非严格约束下的离线策略评估

许多方法都存在相对严格的约束条件. 例如非平稳马尔科夫下的离线策略评估中，通常不考虑时间因素，并假设环境是平稳的，即 MDP 的状态转移和回报函数在每轮迭代之间不会发生变化. 而在现实世界中，强化学习环境通常是非平稳的，因此对应的离线策略评估方法的准确性就可能无法保证. 虽然已经有针对非平稳马尔科夫下的离线策略评估的研究[56-58]，但大多只针对于特定领域的状态空间.

(2) 面向轨迹长度的离线策略评估

在重要性采样方法中，重要性采样是推导无偏估计的关键技术，但会使估计的方差随着轨迹长度的增长指数爆炸，形成“轨迹长度灾难”. 已经有部分工作开始关注这个问题[19,59]，比如直接将 IS 应用于平稳状态访问分布，以避免现有方法所面临的爆炸方差. 但是真正做到对轨迹的长短不一以及噪音不敏感，并应用到现实世界各种场景中，仍需进一步研究.

(3) 混合模型的组合有效性研究

混合模型法中的估计器，包括 DR、WDR、MAGIC 等都需要纯模型法(值迭代、模型近似等)来减少普通重要性采样产生的无偏估计的方差. 由于纯模型法都是采用部分轨迹来得到近似模型，因此效果的好坏取决与不同的技巧(trick). 因此，那些针对不同的场景、不同的轨迹数据、不同纯模型法的技巧、与不同的重要性采样模型组合成混合模型，都具有很大的不确定性，亟待大量的实验和理论研究.

(4) 自评价的离线策略评估

传统离线策略评估的目标是对目标策略的状态价值有一个很好的估计，评判标准是所估计的目标策略状态价值与目标策略真实执行的状态价值有一个较小的均方误差. 因此，大多估计器以减少均方误差为损失函数，例如 MAGIC 等. 而 Google 的 PU 学习法，采用了适用于机器人场景的自评估方法，即通过评估每个 Q 函数等效最优策略的真实回报，比较离线策略分数与真实值的正相关性. 这是很好的开创性工作，然而自评价标准与传统标准的关系与结合，仍然亟待深入研究和探索.

7 总 结

离线策略评估利用行为策略收集到的轨迹数据估计目标策略的状态价值，一直是强化学习非常关键的研究问题. 随着强化学习在机器人、自动驾驶等领域的应用驱动，离线策略评估在近年的人工智能顶刊顶会受到学者们的持续关注.

本文系统性地梳理了近二十年的离线策略评估主要方法：纯模型法、重要性采样法、混合模型法和 PU 学习法. 分别阐述了各类方法的机理、方法中模型的细节差异，详细对各类方法及模型进行了机理对比，并通过亲自实验进行了主流离线策略评估模型的程序复现与性能对比. 最后展望了离线策略评估的技术挑战与可能的未来发展方向，希望对其他学者的研究工作有所启发.

总体来看，离线策略评估取得了很多进展，但各种方法都有各自的适用场景和局限性，特别是对历史轨迹的约束相对较强，因此从理论到实际肯定还有一段很长的路要走. 特别值得注意的是，Google 的 PU 方法在机器人应用中的打破约束确实开了一个好头，为我们未来更加大胆地去探索离线策略评估提供了一个启发.

致 谢 感谢中国科学院计算技术研究所智能信息处理重点实验室和中国科学院信息工程研究所二室相关专家提出的宝贵意见！

参 考 文 献

[1] Murphy M P, Saunders A, Moreira C, et al. The LittleDog robot. The International Journal of Robotics Research, 2011, 30(2): 145-149

[2] Holcomb S D, Porter W K, Ault S, et al. Overview on DeepMind and its AlphaGo Zero AI//Proceedings of the International Conference on Big Data. Seattle, USA, 2018: 67-71

[3] Murphy S A, van der Laan M J, Robins J M, et al. Marginal mean models for dynamic regimes. Journal of the American Statistical Association, 2001, 96(456): 1410-1423

[4] Petersen M, Schwab J, Gruber S, et al. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. Journal of Causal Inference, 2014, 2(2): 147-185

[5] Theocharous G, Thomas P S, Ghavamzadeh M, et al. Personalized ad recommendation systems for life-time value optimization with guarantees//Proceedings of the 24th International Conference on Artificial Intelligence. Amsterdam, Netherlands, 2015: 1806-1812

[6] Hoiles W, Der Schaar M V. Bounded off-policy evaluation with missing data for course recommendation and curriculum design//Proceedings of the 33rd International Conference on Machine Learning. New York, USA, 2016: 1596-1604

[7] Chu W, Li L, Reyzin L, et al. Contextual bandits with linear payoff functions//Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Florida, USA, 2011: 208-214

[8] Langford J, Zhang T. The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information//Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS). Vancouver, Canada, 2007: 817-824

[9] Mannor S, Simester D, Sun P, et al. Bias and variance approximation in value function estimates. Management Science, 2007, 53(2): 308-322

[10] Precup D, Sutton R S, Singh S, et al. Eligibility traces for off-policy policy evaluation//Proceedings of the International Conference on Machine Learning. Stanford, USA, 2000: 759-766

[11] Precup D. Temporal Abstraction in Reinforcement Learning [Ph. D. dissertation]. University of Massachusetts, USA, 2000: 4569-4574

[12] Farajtabar M, Chow Y, Ghavamzadeh M, et al. More robust doubly robust off-policy evaluation//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 1446-1455

[13] Jiang N, Li L. Doubly robust off-policy value evaluation for reinforcement learning//Proceedings of the International Conference on Machine Learning. New York, USA, 2016: 652-661

[14] Thomas P, Brunskill E. Data-efficient off-policy policy evaluation for reinforcement learning//Proceedings of the International Conference on Machine Learning. New York, USA, 2016: 2139-2148

[15] Irpan A, Rao K, Bousmalis K, et al. Off-policy evaluation via off-policy classification//Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS). Vancouver, Canada, 2019: 5437-5448

[16] Littman M L, Dean T, Kaelbling L P, et al. On the complexity of solving Markov decision problems//Uncertainty in Artificial Intelligence. Montreal, Canada, 1995: 394-402

[17] Jong N K, Stone P. Model-based function approximation in reinforcement learning//Adaptive Agents and Multi-Agents Systems. Maastricht, Netherlands, 2007: 1-8

[18] Hanna J P, Niekum S, Stone P, et al. Importance sampling policy evaluation with an estimated behavior policy//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 2605-2613

[19] Xie T, Ma Y, Wang Y X. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling//Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS). Vancouver, Canada, 2019: 9668-9678

[20] Guo Z D, Thomas P S, Brunskill E, et al. Using options and covariance testing for long horizon off-policy policy valuation//Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS). Long Beach, USA, 2017: 2492-2501

[21] Bibaut A F, Malenica I, Vlassis N, et al. More efficient off-policy evaluation through regularized targeted learning//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 654-663

[22] Liu Y, Gottesman O, Raghu A, et al. Representation balancing MDPs for off-policy policy evaluation//Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS). Montréal, Canada, 2018: 2644-2653

[23] Grunewalder S, Lever G, Baldassarre L, et al. Modelling transition dynamics in MDPs with RKHS embeddings//Proceedings of the International Conference on Machine Learning. Edinburgh, UK, 2012: 535-542

[24] Sutton R S, Barto A G. Introduction to Reinforcement Learning. 2nd Edition. Cambridge, UK: MIT Press, 1998

[25] Dann C, Neumann G, Peters J. Policy evaluation with temporal differences: A survey and comparison. Journal of Machine Learning Research, 2014, 15: 809-883

[26] Farahmand A, Szepesvari C. Model selection in reinforcement learning. Machine Learning, 2011, 85(3): 299-332

[27] Marivate V N. Improved Empirical Methods in Reinforcement-Learning Evaluation [Ph. D. dissertation]. Rutgers University New Brunswick, New Jersey, USA, 2015

[28] Jiang N, Kulesza A, Singh S, et al. Abstraction selection in model-based reinforcement learning//Proceedings of the International Conference on Machine Learning. Lille, France, 2015: 179-188

[29] Rubinstein R Y, Kroese D P. Simulation and the Monte Carlo Method. 2nd Edition. New Jersey, USA: John Wiley & Sons, 2016

[30] 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

[31] Cassel C M, Sarndal C E, Wretman J, et al. Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika, 1976, 63(3): 615-620

[32] Robins J M, Rotnitzky A, Zhao L P, et al. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association, 1994, 89(427): 846-866

[33] Robins J M, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association, 1995, 90(429): 122-129

[34] Lunceford J K, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. Statistics in Medicine, 2004, 23(19): 2937-2960

[35] Kang J, Schafer J L. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science, 2007, 22(4): 523-539

[36] Der Laan M J, Rubin D B. Targeted maximum likelihood learning. The International Journal of Biostatistics, 2006, 2(1): 1-40

[37] van der Laan M J, Rose S. Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies. Switzerland: Springer International Publishing, 2018: 35-47

[38] Kiryo R, Niu G, Du Plessis M C, et al. Positive-unlabeled learning with non-negative risk estimator//Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS). Long Beach, California, USA, 2017: 1675-1685

[39] Nagelkerke N J. A note on a general definition of the coefficient of determination. Biometrika, 1991, 78(3): 691-692

[40] Spearman C. The proof and measurement of association between two things. The American Journal of Psychology, 1904, 15(1): 72-101

[41] Cobbe K, Klimov O, Hesse C, et al. Quantifying generalization in reinforcement learning//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 1282-1289

[42] Nichol A, Pfau V, Hesse C, et al. Gotta learn fast: A new benchmark for generalization in RL. arXiv preprint arXiv: 1804.03720, 2018

[43] Raghu M, Irpan A, Andreas J, et al. Can deep reinforcement learning solve erdos-selfridge-spencer games?//Proceedings of the International Conference on Machine Learning. Stockholm, Sweden, 2018: 4235-4243

[44] Zhang A, Ballas N, Pineau J. A dissection of overfitting and generalization in continuous reinforcement learning. arXiv preprint arXiv:1806.03937, 2018

[45] Zhang C, Vinyals O, Munos R, et al. A study on overfitting in deep reinforcement learning. arXiv preprint arXiv: 1804.06893, 2018

[46] Komorowski M, Celi L A, Badawi O, et al. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nature Medicine, 2018, 24: 1716-1720

[47] Bothe M K, Dickens L, Reichel K, et al. The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. Expert Review of Medical Devices, 2013, 10(5): 661-673

[48] Li L, Albertsmet I, Faisal A A, et al. Optimizing medical treatment for sepsis in intensive care: From reinforcement learning to pre-trial evaluation. arXiv preprint arXiv: 2003.06474, 2020

[49] Dudik M, Erhan D, Langford J, et al. Doubly robust policy evaluation and optimization. Statistical Science, 2014, 29(4): 485-511

[50] Swaminathan A, Krishnamurthy A, Agarwal A, et al. Off-policy evaluation for slate recommendation//Proceedings of the Conference and Workshop on the Neural Information Processing Systems (NeurIPS). Long Beach, USA, 2017: 3632-3642

[51] Zou L, Xia L, Ding Z, et al. Reinforcement learning to optimize long-term user engagement in recommender systems//Proceedings of the Knowledge Discovery and Data Mining Anchorage. Alaska, USA, 2019: 2810-2818

[52] Prasad N. Methods for Reinforcement Learning in Clinical Decision Support [Ph. D. dissertation]. Princeton University, Princeton, USA, 2020: 1-7

[53] Futoma J, Masood M A, Doshivelez F, et al. Identifying distinct, effective treatments for acute hypotension with SODA-RL: Safely optimized diverse accurate reinforcement learning. arXiv preprint arXiv:2001.03224, 2020

[54] Li J, Modares H, Chai T, et al. Off-policy reinforcement learning for synchronization in multiagent graphical games. IEEE Transactions on Neural Networks, 2017, 28(10): 2434-2445

[55] Thomas P S. Safe Reinforcement Learning[Ph. D. dissertation]. University of Massachusetts Amherst, Massachusetts, USA, 2015: 5-21

[56] Thomas P S, Theocharous G, Ghavamzadeh M, et al. Predictive off-policy policy evaluation for nonstationary decision problems, with applications to digital marketing//National Conference on Artificial Intelligence(AAAI). San Francisco, USA, 2017: 4740-4745

[57] Dudik M, Erhan D, Langford J, et al. Sample-efficient nonstationary policy evaluation for contextual bandit. arXiv preprint arXiv:1210.4862, 2012

[58] Jagerman R, Markov I, De Rijke M, et al. When people change their mind: Off-policy evaluation in non-stationary recommendation environments//Web Search and Data Mining. Melbourne, Australia, 2019: 447-455

[59] Liu Q, Li L, Tang Z, et al. Breaking the curse of Horizon: Infinite-horizon off-policy estimation//Proceedings of the Conference and Workshop on the Neural Information Processing Systems (NeurIPS). Montréal, Canada, 2018: 5356-5366

WANG Shuo-Ru, M. S. candidate. Her main research interest is cyberspace security.

NIU Wen-Jia, Ph. D., professor. His main research interests include artificial intelligence security, content security.

TONG En-Dong, Ph. D., lecturer. His main research interests include artificial intelligence security, cyber security.

CHEN Tong, Ph. D. candidate. Her main research interests focus on cyberspace security.

LI He, M. S. candidate. His main research interest is cyberspace security.

TIAN Yun-Zhe, M. S. candidate. His main research interest is cyberspace security.

LIU Ji-Qiang, Ph. D., professor. His main research interests include trusted computing, secret protection, cloud computing security.

HAN Zhen, Ph. D., professor. His main research interests include computer security, artificial intelligence and application, system security.

LI Yi-Dong, Ph. D., professor. His main research interests include privacy protection data analysis.

Background

Among reinforcement learning (RL) applications, the off-policy evaluation (OPE) is used to avoid unexpected threats before a RL deployment, that has promising future applied into the fields of the robot, autonomous driving and so on. No needs to actually execute RL, OPE can estimate target policy's state values through history trajectories collected. Generally, the learning goal of OPE is to minimize the mean squared error of state values between the new estimate and actually executed values from target policy.

With the development of AI and RL, as a key technology of RL, OPE has drawn a lot of interest in top AI journals and conferences the recent years. This paper systematically summarizes the main OPE methods of the latest twenty years, including directed-model based, importance-sampling based, hybrid-model based and PU-learning based methods. We present related theory knowledge of OPE, describe detailed differences of different OPE method's mechanisms and models. After performing complete comparisons of methods and corresponding models, we further implement the programs of mainstream OPE models and compare their performance. Finally, we outlook future challenges and possible developing directions of OPE.

In conclusion, OPE has great progress on its way. However, due to the difference between behavior policy and target policy, and possible reward sparseness of behavior policy in some emerging applications, OPE still has a lot of challenges. Happily noting that, Google's PU-learning method breaks some OPE strict constraints, giving inspiration for future OPE researches and grounding OPE applications. Hope that this paper could encourage more and more studies in the OPE field of RL.