无人系统中离线强化学习的隐蔽数据投毒攻击方法

周雪，苘大鹏，许晨，吕继光，曾凡一，高朝阳，杨武
（哈尔滨工程大学计算机科学与技术学院，黑龙江 哈尔滨 150000）

摘　要：针对现有离线强化学习数据投毒攻击方法有效性及隐蔽性不足的问题，提出一种关键时间步动态投毒攻击方法，通过对重要性较高的样本进行动态扰动，实现高效隐蔽的攻击效果。具体来说，通过理论分析发现时序差分误差对于模型学习过程具有重要影响，将其作为投毒目标选择的依据；进一步提出基于双目标优化的投毒方法，在最小化扰动幅度的同时，最大化攻击对模型性能产生的负面影响，为每个投毒样本生成最优扰动幅度。在多种任务及算法中的实验结果表明，所提攻击方法仅在投毒比例为整体数据 1% 的情况下，就能使智能体的平均性能下降 84%，揭示了无人系统中离线强化学习模型的敏感性及脆弱性。

关键词：无人系统；离线强化学习；数据投毒攻击；数据安全
中图分类号：TP393
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024264

Stealthy data poisoning attack method on offline reinforcement learning in unmanned systems

ZHOU Xue, MAN Dapeng, XU Chen, LYU Jiguang, ZENG Fanyi, GAO Chaoyang, YANG Wu
College of Computer Science and Technology, Harbin Engineering University, Harbin 150000, China

Abstract: Aiming at the limitations in effectiveness and stealth of existing offline reinforcement learning(RL) data poisoning attacks, a critical time-step dynamic poisoning attack was proposed, perturbing important samples to achieve efficient and covert attacks. Temporal difference errors, identified through theoretical analysis as crucial for model learning, were used to guide poisoning target selection. A bi-objective optimization approach was introduced to minimize perturbation magnitude while maximizing the negative impact on performance. Experimental results show that with only a 1% poisoning rate, the method reduces agent performance by 84%, revealing the sensitivity and vulnerability of offline RL models in unmanned systems.

Keywords: unmanned system, offline reinforcement learning, data poisoning attack, data security

0　引言

在无人系统的智能决策与控制中，强化学习（RL, reinforcement learning）技术发挥着关键作用。通过智能体与环境的交互学习，不断优化行为策略，使无人系统能够自主实现高效决策[1-3]。然而，传统在线强化学习在无人系统中的应用面临试错成

收稿日期：2024–09–02；**修回日期：**2024–11–25
通信作者：许晨，chen.xu@hrbeu.edu.cn
基金项目：国家重点研发计划基金资助项目（No.2021YFB3101401）；黑龙江省自然科学基金资助项目（No.TD2022F001）；国家自然科学基金资助项目（No.U2003206, No.U20B2048, No.U21B2019, No.U22A2036, No.62272127）；中央高校基本科研业务费专项资金资助项目（No.3072024XX0607）
Foundation Items: The National Key Research and Development Program of China (No.2021YFB3101401), The Natural Science Foundation of Heilongjiang Province (No.TD2022F001), The National Natural Science Foundation of China (No.U2003206, No.U20B2048, No.U21B2019, No.U22A2036, No.62272127), The Fundamental Research Funds of the Central Universities (No.3072024XX0607)

本高、实时性要求高等挑战[4]。因此，离线强化学习因其无需实时与环境进行交互，仅利用预收集的经验数据进行训练，显著降低了试错成本[5]，近年来受到了广泛关注。离线强化学习在自动驾驶[6]、智能机器人控制[7]等应用中展现出巨大潜力，为实现无人系统的高效能、智能化提供了有力支撑。

虽然离线强化学习在多个任务中展示了其有效性，但其性能在很大程度上依赖于离线数据集的质量[8]，质量不佳的数据很容易导致学习效率低下，甚至产生误导性结果[9]。研究表明，在高风险决策任务中，数据的安全性和可信性对保证无人系统中模型的有效性和可靠性至关重要[10]。数据中毒攻击通过恶意修改训练数据观察算法的学习效果[11]，从而评估其数据敏感性及脆弱性，是目前常用的一种有效方法[12-13]。在离线强化学习中，攻击者可以通过在离线数据集中植入伪造或篡改的数据，使得智能体在训练过程中学习到错误的信息，从而导致其在实际应用中做出错误决策。通过分析数据投毒攻击给智能体带来的影响，可以充分了解智能体在离线数据集遭受攻击时的敏感性及脆弱性，对于提高离线强化学习模型的有效性和可靠性至关重要。

现有针对强化学习的投毒攻击研究主要集中在在线阶段[14-19]，攻击者通过操控奖励信号、篡改环境反馈或直接干预策略更新等方式，显著影响了智能体的决策能力。例如，攻击者通过在训练过程中注入微小扰动，使智能体执行错误策略，甚至完全偏离预期目标[20-22]。此外，一些研究关注强化学习系统中薄弱环节的识别。部分方法将动作偏好值（C值）[23]、动作价值（Q值）[24]和熵值[25]超过阈值的时刻视为关键时间步。然而，在复杂任务中，高价值时间步未必总是关键的。例如，在自动驾驶场景中，具有高价值的时间步通常对应于直行场景，但根据以往经验，转弯的时间步通常更为关键。为此，Sun等[26]通过测量汽车与道路中线的距离识别关键时间步。此外，Sun等[21]还考虑了学习理论中的稳定性概念，以确定关键时间步。Yu等[27]训练了专门的模型来确定关键时间步，但这种特定模型在应用于新环境时经常表现出较差的泛化能力。这些在线攻击方式揭示了强化学习系统在面对恶意攻击时的脆弱性，也为后续研究离线强化学习的数据安全提供了宝贵经验。然而，这些攻击方法需要实时获取强化学习的在线训练参数等信息，因此无法直接适用于缺乏实时交互的离线场景。

目前有少部分针对离线强化学习数据投毒攻击的研究。Ma等[28]首次对离线强化学习数据集进行了投毒攻击。然而，该方法针对每个时间步进行攻击，投毒比例过高易被检测。Rakhsha等[29]提出了平衡效果和代价的攻击方法，但仅关注了简单的离散型任务，没有针对更贴近真实场景的连续型任务（如机器人、自动驾驶等）。Gong等[30]提出了在离线数据集中插入触发器的攻击方法，但需要攻击者在训练和测试阶段都能够执行攻击，对攻击者的权限要求过高。上述方法存在投毒比例较高、扰动幅度较大导致攻击有效性不足且不够隐蔽的问题，难以模拟现实世界中隐蔽性攻击的场景，限制了离线强化学习防御策略的研究。分析认为，上述工作主要存在以下问题。

1) 未选择合适投毒目标。离线数据集中每个样本对策略学习的贡献不同，而现有方法未能精准识别并选择具有关键影响的目标样本进行投毒，导致攻击代价较大且有效性不足。

2) 未选择最优扰动幅度。现有工作中扰动幅度过高，易产生显著异常数据，导致攻击不够隐蔽。且不同样本对应的最优扰动幅度存在差异，而现有方法往往采用固定幅度，导致攻击缺乏有效性。

针对上述问题，本文提出一种关键时间步动态投毒攻击（CTDPA, critical time-step dynamic poisoning attack）方法。首先从离线强化学习理论层面分析得出，时序差分（TD, temporal difference）误差较大的时间步能够严重影响模型学习效果，并依据TD误差值定位关键时间步。其次，提出一种双目标优化的投毒攻击（DOPA, dual-objective poisoning attack），将扰动幅度选择形式化为最小化约束扰动同时最大化TD误差的双目标优化问题，针对每个关键样本求解出更加隐蔽且使模型性能显著下降的动态扰动。本文主要的贡献包括以下3个方面。

1) 针对离线强化学习投毒攻击比例较高且扰动幅度较大的问题，提出一种关键时间步动态投毒攻击方法，在保证隐蔽的前提下提升攻击有效性。

2) 基于TD误差定位关键时间步，通过攻击学习过程中的关键环节，提升攻击效率。基于双目标

优化方法选择最优扰动幅度，实现在微小扰动下显著降低智能体性能。

3) 在4个连续型复杂任务的数据集和4种主流离线强化学习算法上进行攻击实验，结果表明本文方法仅在投毒比例为干净数据的1%时，能够使智能体的平均性能下降84%。且当攻击者仅有小范围数据权限时，也具有一定的攻击效果。

1 理论基础

1.1 离线强化学习

离线强化学习是强化学习的一种范式，以下是离线强化学习的定义：在一个离线强化学习任务中，考虑一个马尔可夫决策过程（MDP, markov decision process），可以将MDP描述为一个4元组$(\mathcal{S},\mathcal{A},\mathcal{R},\mathcal{P})$，由状态空间$\mathcal{S}$、动作空间$\mathcal{A}$、奖励函数$\mathcal{R}$和状态转移概率$\mathcal{P}$组成。离线强化学习智能体从预先收集好的数据集中学习并改进其策略π。预先收集的数据集表示为D，数据集D中的轨迹是通过某个已有的策略生成，t时刻的轨迹$\tau=(s_t,a_t,r_t,s_{t+1})$。其中$s_t$是当前状态，$a_t$为智能体在状态$s_t$处根据策略$\pi$选择的动作，接收到的奖励$r_t$由奖励函数$\mathcal{R}$给出，即$r_t=\mathcal{R}(s_t,a_t)$。同时状态根据状态转移概率$\mathcal{P}$由$s_t$转移到下一个状态$s_{t+1}$，即$s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)$。离线强化学习的目标是从离线数据集$D$中学习得到一个能够获得最大累积奖励的策略$\pi^*$，可以描述为如下的优化问题。

$$\pi^*=\arg\max_{\pi}\mathbb{E}_{\tau\sim D}[R(\tau^{\pi})] \tag{1}$$

其中，$R(\tau)=\sum_{t=0}^{T}\gamma^t r_t$为累积奖励，$T$为轨迹数量，$\gamma$是折扣因子[31]，表示当前奖励对未来奖励的重要程度，$\mathbb{E}$表示期望。

1.2 威胁模型

离线强化学习的应用思想是数据提供者可以以开源的方式共享自己的经验数据，开发者可以利用开源数据训练强化学习智能体来减少消耗，任何人都可以成为开源数据的上传者。参考现有研究的设置[28-30]，本文针对离线强化学习的威胁模型如图1所示。

1) 攻击者权限及知识

将攻击者设定为具有数据操作权限的人员，可以是恶意数据提供者，也可以是恶意数据处理人员。攻击者有能力对离线数据集D中的部分数据进行投毒，生成中毒数据集D'，但攻击者为了实施更加隐蔽的攻击，需要尽可能限制投毒比例和扰动幅度的大小。数据集D的数据总量为N，N_p表示中毒数据的数量，投毒比例定义为$p=\frac{N_p}{N},p\in[0,1]$。为了最小化攻击代价，本文中毒数据数量远小于数据总量，可以表示为$N_p\ll N$。

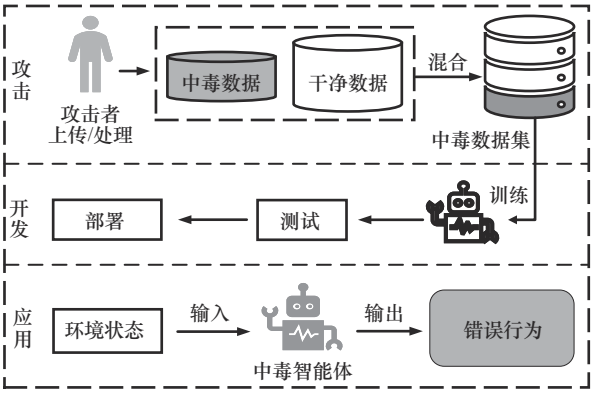

图1 威胁模型

攻击者不需要任何领域知识（如无人机器人中动力学与感知、自动驾驶中计算机视觉与车辆动力学等），且攻击者无需了解具体的训练任务。

2) 攻击目标

与正常离线强化学习过程相反，攻击者的目标是使智能体通过学习中毒数据集D'，获得能够最小化累积奖励，甚至获得惩罚的策略π'，使智能体根据学习到的中毒策略π'做出错误动作，可以描述为

$$\pi'=\arg\min_{\pi}\mathbb{E}_{\tau\sim D'}[R(\tau^{\pi})] \tag{2}$$

3) 攻击指标

现有针对强化学习数据投毒的研究中[30]，通常通过中毒模型相较干净模型的平均累积奖励下降程度，即性能下降比例（PDR, performance decline rate）来衡量攻击的有效性。智能体在测试环境中的平均累积奖励为

$$\bar{R}=\frac{1}{T}\left[\sum_{t=0}^{T}r_t\right] \tag{3}$$

其中，T为轨迹数量。与策略更新过程不同的是，$\bar{R}$计算时不包含折扣因子γ。干净智能体在测试环境中获得的平均累积奖励表示为$\bar{R}_{\text{clean}}$，中毒智能体获得的平均累积奖励表示为$\bar{R}_{\text{poisoned}}$，中毒智能体PDR表示为

$$\mathrm{PDR}=\frac{\bar{R}_{\mathrm{clean}}-\bar{R}_{\mathrm{poisoned}}}{\bar{R}_{\mathrm{clean}}}\times 100\% \tag{4}$$

中毒智能体性能下降越多表明攻击效果越好。隐蔽攻击旨在用最小的投毒比例p，使中毒智能体性能显著下降。

2　方法设计

为了模拟潜在的数据缺陷，从而揭示精心设计的攻击对离线强化学习算法的影响，本文提出了一种关键时间步动态投毒攻击方法，将攻击过程分解为2个部分：关键时间步定位和双目标优化投毒攻击，其方法架构如图2所示。具体来说，首先识别数据集中在受到攻击时会对智能体性能产生更大影响的时间步，缩小攻击范围，降低投毒代价从而提升攻击效率。其次，将扰动幅度选择转化为双目标优化问题，通过最小化扰动幅度、最大化攻击影响，以实现更加隐蔽的攻击。

2.1　基于TD误差的关键时间步定位方法

现有工作的关键时间步定位方法均为在线强化学习设计，依靠与环境的实时交互信息来定位关键时间步。然而，在离线强化学习中，攻击者无法获取这些关键信息，这使得上述方法无法直接应用于离线阶段。

为解决上述问题，本文提出一种具有理论依据的关键时间步定位方法。首先对离线强化学习的更新过程进行了理论分析，推断出TD误差较大的时间步对于模型学习影响更为显著。其次，通过训练正常智能体来计算每个时间步的TD误差，获得关键时间步集合用于后续投毒攻击。

在离线强化学习中，$R(\tau)=\sum_{t=0}^{T}\gamma^{t}r_t$表示在策略$\pi$下获得的累积折扣奖励，在特定状态$s$遵循策略$\pi$执行动作$a$后的累积折扣奖励期望值为状态-动作价值，表示为$Q^{\pi}(s,a)$。

$$Q^{\pi}(s,a)=\mathbb{E}_{\pi}[R(\tau^{\pi})|s_0=s,a_0=a] \tag{5}$$

为了提高更新过程的效率，现有离线强化学习算法大都采用时序差分算法来量化当前时间步和下一时间步间估计值的差异，从而使每个时间步都能进行充分学习。参照Q-Learning算法[32]，更新机制可表述为

$$\begin{aligned}&Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha\delta_t\\&\delta_t=r_t+\gamma\max_{a+1}Q(s_{t+1},a_{t+1})-Q(s_t,a_t)\end{aligned} \tag{6}$$

其中，δ_t表示TD误差，$\max_{a}Q(s_{t+1},a_{t+1})$表示后续状态$s_{t+1}$中所有潜在行动最大$Q$值的期望值，$\alpha$是学习率，$\gamma$是折扣因子。现有研究（如Schaul等[33]）表明，具有较大δ_t的样本表现出更高的学习价值，表明智能体对t时刻输入数据的预测准确性不足，为智能体学习过程中的薄弱环节，从这些时间步中进一步学习可以帮助策略改进。

因此，基于上述分析结果，本文有针对性地将具有较大TD误差δ_t的时间步作为关键时间步，并对其进行投毒攻击，从而在降低攻击代价的同时对学习过程产生更大的损害。基于TD误差的关键时间步定位方法具体通过以下3个步骤完成。

1) 干净智能体训练：本文的攻击者为恶意数据上传方或具有数据处理权限的人员，不具备获取训练参数的权限，因此采用一种间接的策略，即在本地使用干净数据集预训练一个具有正常水平的智能体A，训练过程可以采用与目标任务不同的算法。首先，基于离线强化学习算法，使用干净数据集D对智能体进行训练，并在测试环境中对训练模型进行性能评估，根据评估结果调整训练参数，使其获得的平均累积奖励$\bar{R}_{\mathrm{clean}}$能够达到现有研究中

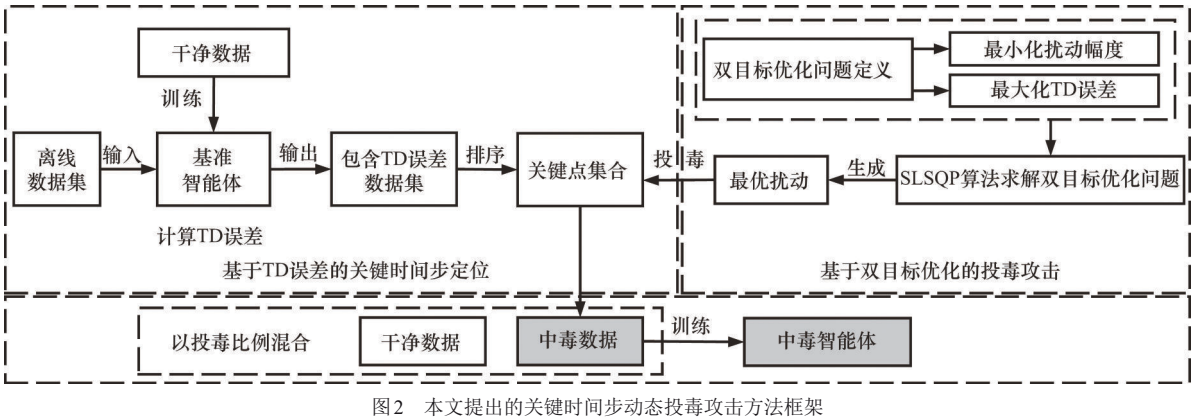

图2　本文提出的关键时间步动态投毒攻击方法框架

的平均水平。

2) TD 误差计算：获取上述干净智能体的动作价值 Q 网络参数 θ，通过 Q 函数 $Q_\theta(s,a)$ 对 t 时刻的状态−动作对 (s_t,a_t) 计算 $Q_\theta(s_t,a_t)$，获取 $t+1$ 时刻的状态−动作对 (s_{t+1},a_{t+1}) 并计算 $Q_\theta(s_{t+1},a_{t+1})$。之后，根据式(6)计算数据集中每个时间步的 TD 误差 δ_t。

3) 确定关键时间步集合：将数据集根据 TD 误差 δ_t 由大到小进行排序，按照投毒比例 p 选择 δ_t 较大的时间步作为关键时间步，并得到关键时间步集合 C 作为后续数据投毒的攻击对象。

基于 TD 误差的关键时间步定位方法如算法 1 所示。

算法 1 基于 TD 误差的关键时间步定位方法

输入 数据集 D、数据集 D 的数量 N、投毒比例 p、学习率 α、折扣因子 γ

输出 关键时间步集合 C

1)初始关键时间步集合 $C \leftarrow \{\}$

2)在本地环境使用干净数据集 D 训练干净智能体 A，得到动作价值函数 $Q_\theta(s,a)$

3)计算 TD 误差集合 $E \leftarrow \{\}$

4)for 每一个时间步 $t \in D$ do

5) 获取当前时间步的状态 s_t，动作 a_t，奖励 r_t，下一时间步状态 s_{t+1}，下一时间步动作 a_{t+1}

6) 利用动作价值函数根据式(6)计算当前时间步的 TD 误差值 δ_t

7) 将 δ_t 作为新的一列添加到原始数据集 D 中 t 时刻对应位置

8)end for

9)对数据集 D 根据 TD 误差 δ_t 列进行降序排序

10)选取前 $\lfloor pN \rfloor$ 个时间步的索引集合 ID

11)for 每一个索引 $i \in$ ID do

12) 获取当前索引对应的数据

13) 将数据添加到关键时间步集合 C 中

14)end for

15)返回关键时间步集合 C

2.2 基于双目标优化的投毒攻击方法

为了实现隐蔽投毒攻击，常用方法是对添加扰动的幅度进行限制[34]。现有针对在线强化学习的攻击研究中也有类似的设置[35]。但这些方法采用的扰动幅度仍然较大，可能会导致显著性数据异常，使得攻击不够隐蔽。同时可能会有部分扰动虽然幅度较大但对模型的影响较小，导致攻击有效性不足。

为解决上述问题，本文提出一种 DOPA 方法。现有研究表明，在离线强化学习中奖励信号的作用可能不如其他信号显著[36-37]，因此本文将状态−动作对 (s,a) 作为攻击对象。在确定关键时间步后，对关键时间步集合 C 中的 (s,a) 添加扰动 η，影响智能体对 Q 值估计的准确性，进而显著增大 TD 误差。为了在不引起明显异常的情况下显著降低智能体性能，达到隐蔽攻击的目的，将扰动幅度的选择形式化为带约束双目标优化问题，具体通过以下 3 个步骤完成。

1) 双目标优化问题定义：优化目标是针对关键时间步集合 C 中每个时间步 t 对应的状态−动作对 (s_t,a_t)，找到最优的扰动 $\eta_{(s_t,a_t)}$，该扰动同时满足 2 个优化目标：一是扰动幅度最小化以保持攻击隐蔽性；二是使 TD 误差 δ_t 最大化以增大攻击产生的影响。可以形式化为式(7)所示的带约束的双目标优化问题。

$$\begin{aligned} &\min_{\eta_{(s_t,a_t)}} \ \|\eta_{(s_t,a_t)}\|_2 \\ &\max_{\eta_{(s_t,a_t)}} \ \delta_t' \\ &\text{s.t. } \|\eta_{(s_t,a_t)}\|_2 \leqslant \varepsilon \end{aligned} \tag{7}$$

其中，$\|\eta_{(s_t,a_t)}\|_2$ 表示对 t 时间步状态−动作对添加扰动 $\eta_{(s_t,a_t)}$ 的 L2 范数，用于量化扰动幅度大小。$\delta_t' = r_t + \gamma \max_{a+1} Q(s_{t+1},a_{t+1}) - Q((s_t,a_t) + \eta_{(s_t,a_t)})$ 为添加扰动后的 TD 误差。ε 是扰动幅度的上限，用于限制扰动幅度的大小。

2) 双目标优化问题求解：上述优化问题中将扰动 $\eta_{(s_t,a_t)}$ 限制在 ε 范围内属于边界约束问题，在优化求解过程中需要确保扰动大小不超过约束范围。TD 误差 δ_t 计算过程中的值函数 $Q(s,a)$ 预测与折扣因子 γ 计算涉及非线性函数，属于非线性约束。为便于求解，采用加权和法将上述双目标优化问题转化为单目标优化问题。

$$\begin{aligned} &\min_{\eta_{(s_t,a_t)}} \ (\beta_1\|\eta_{(s_t,a_t)}\|_2 - \beta_2\delta_t') \\ &\text{s.t. } \ \|\eta_{(s_t,a_t)}\|_2 \leqslant \varepsilon \end{aligned} \tag{8}$$

其中，β_1 和 β_2 是权重参数，用于平衡 2 个目标的重要性，本文中 2 个目标的重要程度相同，因此 β_1 和

β_2均取1。为解决上述优化问题，本文采用序列最小二乘规划（SLSQP, sequential least squares programming）算法[38]作为优化求解器对目标函数$L(\eta_{(s_t,a_t)})$进行求解。SLSQP算法是一种用于解决非线性约束优化问题的迭代方法，能够高效地求解包含等式和不等式约束的优化问题。

3) 投毒攻击实施：将求解得到的最优扰动$\eta^*_{(s_t,a_t)}$添加到原始数据集D中关键时间步对应的状态−动作对(s_t,a_t)上，生成中毒数据集D'。用数据集D'对智能体进行训练，得到中毒智能体A'。在测试环境中评估中毒智能体A'性能相对干净智能体A的下降比例，进而量化攻击对离线强化学习智能体产生的负面影响。

基于双目标优化的投毒攻击方法如算法2所示。

算法2　基于双目标优化的投毒攻击方法

输入　数据集D、关键时间步集合C、扰动幅度上限ε、学习率α、折扣因子γ、优化求解器SLSQP算法

输出　中毒数据集D'、中毒智能体A'

1)初始中毒数据集$D' \leftarrow D$

2)for 每一个时间步 $t \in C$ do:

3)　初始扰动$\eta_{(s_t,a_t)} \leftarrow 0$

4)　获取当前时间步的状态s_t，动作a_t，奖励r_t，下一时间步状态s_{t+1}，下一时间步动作a_{t+1}

5)　定义优化目标：最小化扰动$\eta_{(s_t,a_t)}$并最大化TD误差δ_t

6)　将优化问题形式化为：$\min\limits_{\eta_{(s_t,a_t)}} \|\eta_{(s_t,a_t)}\|_2$，$\max\limits_{\eta_{(s_t,a_t)}} \delta'_t$，且约束条件为$\|\eta_{(s_t,a_t)}\|_2 \leqslant \varepsilon$

7)　为便于求解，采用加权和法将双目标优化问题转化为单目标优化问题

$\min\limits_{\eta_{(s_t,a_t)}} (\beta_1\|\eta_{(s_t,a_t)}\|_2 - \beta_2\delta'_t)$且$\|\eta_{(s_t,a_t)}\|_2 \leqslant \varepsilon$

8)　利用动作价值函数$Q(s,a)$根据式(6)计算当前时间步的TD误差值δ_t

9)　使用SLSQP算法求解优化问题，得到最优扰动$\eta^*_{(s_t,a_t)}$

10)　更新当前时间步的状态−动作对$(s'_t,a'_t) \leftarrow (s_t,a_t) + \eta^*_{(s_t,a_t)}$

11)　将扰动后的状态−动作对(s'_t,a'_t)更新到中毒数据集D'中对应的位置

12)end for

13)使用中毒数据集D'训练智能体，得到中毒智能体A'

14)返回中毒数据集D'和中毒智能体A'

3　实验分析

3.1　实验环境

实验的服务器配置为：Intel(R) Xeon(R) Silver 4214R CPU @ 2.40 GHz，128 GB内存，NVIDIA GeForce RTX 3090显卡（24 GB显存）。使用的编程语言和主要开源软件库版本为Python 3.8、PyTorch 1.6.0、D4RL 1.1。

3.2　数据集和算法

1) 数据集：为了全面评估本文提出的攻击方法，在Fu等[39]提出的离线环境D4RL中选择了4个复杂的连续型任务：MuJoCo机器人任务[40]中的Hopper、Half-Cheetah和Walker2D，3个任务对应数据集均为中等水平；Carla自动驾驶任务[41]中的Carla-Lane，这些任务更接近真实世界场景。MuJoCo机器人任务要求智能体以不同形式快速前进，Carla-Lane自动驾驶任务要求智能体平稳、快速地驾驶汽车前进，数据集的具体信息如表1所示。

表1　本文使用的数据集情况

环境	任务类型	任务	数据集	数据量
MuJoCo	机器人控制	Hopper	hopper-medium	2×10^6
		Half-Cheetah	halfcheetah-medium	10^6
		Walker2D	walker2d-medium	10^6
Carla	自动驾驶	Carla-Lane	carla-lane	10^5

2) 离线强化学习算法：为了评估不同离线强化学习算法在本文攻击方法下的鲁棒性，采用4种算法来训练智能体：批约束深度Q学习（BCQ, batch-constrained Q-learning）算法[42]、回溯批量演员−评论家（BEAR, batch-ensemble actor-critic with retrace）算法[43]、保守Q学习（CQL, conservative Q-Learning）算法[44]和行为克隆（BC, behavioural cloning）算法[31]。训练过程使用上述算法的官方开源代码，并遵循D4RL基准工作中的参数设置[39]。

3.3　实验设置

尽管面向在线强化学习的攻击方法依赖实时获取训练过程中的参数，无法直接适用于离线场景，

但为更全面地衡量本文方法的性能，本文首先在上述4种环境中将CTDPA方法与现有数据投毒攻击方法进行对比，具体包括面向在线强化学习的鲁棒Sarsa（RS）方法、状态对抗强化学习（SA-RL）方法[15]和策略对抗演员导演（PA-AD）方法[18]，以及面向离线强化学习的BAFFLE方法[30]。由于上述在线攻击方法未在Carla环境中进行实验，因此本文仅与离线攻击方法进行对比。在线强化学习采用近端策略优化（PPO, proximal policy optimization）算法[45]训练智能体，离线强化学习采用CQL算法。攻击性能通过中毒智能体相较干净智能体的平均性能下降比例PDR来衡量，性能下降越大表明攻击对目标策略的破坏性越强。

其次，为分析离线强化学习算法在面对精心设计的攻击时的敏感性及脆弱性，本文进行了攻击方法的性能实验，在4个数据集对应4种离线强化学习算法中评估攻击方法性能。主要通过2个关键环节来实现隐蔽攻击，一是通过定位关键时间步缩小投毒范围，以降低攻击代价进而实现更加有效的攻击；二是基于双目标优化思想针对关键时间步对应数据生成中毒扰动，在最小化扰动大小的同时最大化攻击产生影响，以提升攻击的隐蔽性及有效性。

之后，为了验证2个关键环节对提升攻击性能的作用，本文进行了消融分析实验。将提出的基于TD误差的关键时间步定位方法与随机投毒以及2种在线场景中确定关键时间步方法进行了对比。并将本文提出的约束内基于双目标优化的投毒攻击方法与现有在线场景中直接添加约束内扰动、固定扰动幅度的方法进行了对比，进一步验证了2个关键环节对攻击性能的作用。

最后，对投毒比例p进行了敏感性实验及结果分析，探究了p的大小对攻击性能的影响。

3.4 对比实验与结果分析

表2展示了本文提出的CTDPA方法与现有攻击方法在4种环境中的攻击有效性对比实验结果。表2结果表明，RS、SA-RL和PA-AD这3种面向在线强化学习攻击方法均能够取得一定的攻击效果，但投毒比例需达到50%，扰动幅度也较高。在相同设置下CTDPA方法能够使智能体性能下降更多，且CTDPA方法在实现相近攻击效果时需要的代价更小。例如，在Hopper环境中，CTDPA方法仅需要在线方法$\frac{1}{5}$的投毒比例（即p=10%）就可达到相似的攻击效果。一方面可能由于在线强化学习在训练过程中不断更新策略并通过与环境交互修正偏差，使得中毒数据对学习过程产生的影响在一定程度上被抑制，进而需要更高的投毒比例才能实现显著的攻击效果。而离线强化学习在训练过程中无需与环境实时交互，因此中毒数据对策略的影响更加显著。另一方面CTDPA方法首先精确定位对智能体性能影响较大的关键时间步，相比RS方法的随机扰动和SA-RL方法、PA-AD方法的全局攻击，缩小了攻击范围，提升攻击效率。同时，将扰动生成转化为双目标优化问题，相比上述方法中的攻击策略，能够在更隐蔽的情况下实现高效攻击。

表2 与现有工作方法的攻击有效性对比实验结果

环境	攻击方法	强化学习	p	ε	性能下降比例
Hopper	RS	在线	50%	0.075	75%
	SA-RL	在线	50%	0.075	80%
	PA-AD	在线	50%	0.075	95%
	CTDPA	离线	50%	0.075	99%
	BAFFLE	离线	10%	—	37%
	CTDPA	离线	10%	0.050	95%
	CTDPA	离线	5%	0.050	94%
Half-Cheetah	RS	在线	50%	0.150	93%
	SA-RL	在线	50%	0.150	109%
	PA-AD	在线	50%	0.150	105%
	CTDPA	离线	50%	0.150	112%
	BAFFLE	离线	10%	—	64%
	CTDPA	离线	10%	0.050	91%
	CTDPA	离线	5%	0.050	88%
Walker2D	RS	在线	50%	0.050	70%
	SA-RL	在线	50%	0.050	76%
	PA-AD	在线	50%	0.050	82%
	CTDPA	离线	50%	0.050	98%
	BAFFLE	离线	10%	—	53%
	CTDPA	离线	10%	0.050	94%
	CTDPA	离线	5%	0.050	93%
Carla-Lane	BAFFLE	离线	10%	—	48%
	CTDPA	离线	10%	0.050	86%
	CTDPA	离线	5%	0.050	84%

与面向离线强化学习的BAFFLE方法相比，当投毒比例相同（即p=10%）时，CTDPA方法在4个环境中的性能下降比例分别提升了58%、27%、41%和38%，平均提升了41%。当CTDPA方法的投毒比例仅为BAFFLE方法的$\frac{1}{2}$（即p=5%）时，CTDPA方法也优于BAFFLE方法，性能下降比例平均提升了39%。结果表明CTDPA方法在相同甚至更低的投毒代价下，能够显著提高攻击效果。这是由于本文方法将训练过程中的关键时间步作为攻击对象，且针对关键时间步生成动态最大化负面影响的扰动，从而提升攻击性能。

综合来看，本文提出的CTDPA方法能够以较低投毒比例及微小扰动幅度实现较好的攻击效果，表明该方法在高效性与隐蔽性方面具有显著优势。

3.5　攻击性能实验与结果分析

本文从有效性、泛化性及隐蔽性3个方面揭示提出的攻击方法对离线强化学习算法产生的影响。

1) 有效性：表3展示了投毒比例p为1%和5%时的攻击效果，表3中数字为智能体在测试环境中运行50条轨迹的平均累积奖励，括号中为智能体性能下降比例，其中扰动幅度ε取0.05，仅为原始数据大小的5%。结果表明，本文方法能够以较小的投毒比例及扰动幅度，在不同任务中显著破坏离线强化学习算法性能，体现了本文方法攻击的有效性。

2) 泛化性：虽然本文的理论分析过程是基于Q值的更新过程进行，但为了验证方法的泛化性，在基于CQL算法计算得到的中毒数据中训练了BC智能体。从表3可以看出，攻击方法针对未基于Q值更新的BC算法依旧有效。这是由于本文方法本质上是通过定位关键时间步，并放大关键时刻预测值与真实值之间的偏差来实现攻击。尽管BC算法不直接依赖于TD误差，但其学习目标仍为最小化预测值与真实值之间的误差。而本文方法通过投毒有效地破坏了这种最小化过程，影响了模型减少预测误差这一基本学习机制，因此能够在对后续任务和算法未知的情况下，极大程度地损害智能体性能，证明了攻击方法的泛化性。

3) 隐蔽性：为了使攻击更加隐蔽，本文进一步增加了对攻击者权限及能力的限制，假设攻击者只能上传或修改少部分恶意数据，无法对完整数据集进行投毒攻击。在Walker2D数据集中随机抽取连续20%、10%、5%和1%时间步组成的数据段，并分别对这些数据段实施数据投毒攻击。实验结果如表4所示，结果表明即使攻击者具有较小的数据访问或处理权限，也能够通过攻击显著影响智能体性能，这也进一步证明了本文方法的有效性。且攻击效果随着数据权限的增加而增加，以CQL算法为例，攻击性能随攻击者权限大小的变化情况如图3所示。这是由于在较大的数据范围中攻击者能够定位到数据集中更加关键的时间步。

表3　针对4种环境4种算法的攻击有效性实验结果

环境	算法	基线	投毒比例	
			p = 1%	p = 5%
Hopper	BCQ	2 823	293(90%)	223(92%)
	BEAR	2 119	204(90%)	131(94%)
	CQL	3 158	415(87%)	190(94%)
	BC	3 450	378(89%)	256(93%)
	平均	2 613	323(89%)	200(93%)
Half-Cheetah	BCQ	4 694	867(82%)	630(87%)
	BEAR	4 290	531(88%)	394(91%)
	CQL	4 822	652(86%)	562(88%)
	BC	4 017	543(86%)	406(89%)
	平均	4 456	648(86%)	498(89%)
Walker2D	BCQ	2 341	179(92%)	147(94%)
	BEAR	2 593	226(91%)	198(92%)
	CQL	3 132	264(92%)	230(93%)
	BC	744	59 (92%)	34(95%)
	平均	2 203	182(92%)	152(94%)
Carla-Lane	BCQ	466	137(70%)	65(86%)
	BEAR	89	25(71%)	17(81%)
	CQL	191	58(70%)	31(84%)
	BC	384	130(66%)	56(85%)
	平均	283	88(69%)	43(84%)
平均性能下降比例			84%	90%

同时本文还探讨了不同扰动幅度ε上限的大小对攻击效果的影响。观察到随着扰动幅度的增加，攻击使智能体性能的下降比例越大，攻击效果越好。且较大的扰动幅度上限能够弥补数据权限较小时攻击效果不佳的情况。例如当扰动幅度ε上限由0.05增加到0.15，投毒比例为1%时，攻击者仅具有5%的数据权限就能够实现与20%的数据权限相近的攻击效果。

表4　Walker2D环境中不同权限下的攻击实验结果

算法	基线	ε	p = 1%				p = 5%			
			20%数据	10%数据	5%数据	1%数据	20%数据	10%数据	5%数据	1%数据
BCQ	2341	0.05	406(83%)	529(77%)	595(75%)	726(69%)	279(88%)	322(86%)	392(83%)	623(73%)
		0.10	362(85%)	413(82%)	415(82%)	679(71%)	225(90%)	272(88%)	351(85%)	554(76%)
		0.15	324(86%)	388(83%)	385(84%)	647(72%)	288(92%)	252(89%)	316(86%)	519(78%)
BEAR	2593	0.05	429(83%)	554(79%)	668(74%)	802(69%)	302(88%)	394(85%)	499(81%)	696(73%)
		0.10	384(85%)	443(83%)	541(79%)	725(72%)	287(89%)	337(87%)	415(84%)	606(77%)
		0.15	329(87%)	384(85%)	484(81%)	622(76%)	200(92%)	251(90%)	358(86%)	516(80%)
CQL	3132	0.05	449(86%)	510(84%)	656(79%)	920(71%)	315(90%)	378(88%)	468(85%)	796(75%)
		0.10	410(87%)	472(85%)	550(82%)	793(75%)	270(91%)	302(90%)	407(87%)	587(81%)
		0.15	304(90%)	343(89%)	422(87%)	671(79%)	195(94%)	243(92%)	308(90%)	543(83%)
BC	744	0.05	95(87%)	130(83%)	140(81%)	232(69%)	87(88%)	91(88%)	112(85%)	204(73%)
		0.10	87(88%)	113(85%)	127(83%)	193(74%)	72(90%)	81(89%)	100(87%)	171(77%)
		0.15	74(90%)	107(86%)	99(87%)	153(79%)	58(92%)	65(91%)	86(88%)	128(83%)

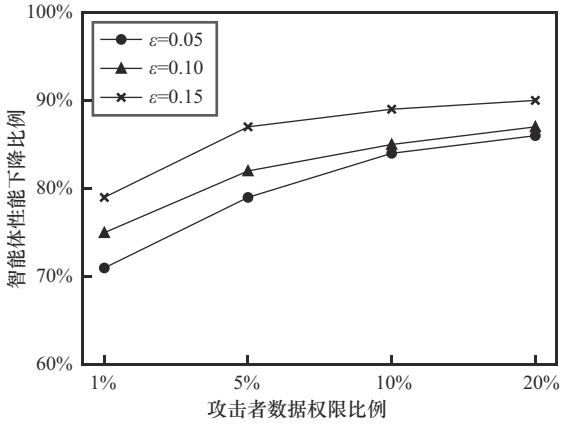

图3　不同数据权限实验结果对比

3.6　消融实验与结果分析

1) 定位关键时间步方法对比。为了讨论不同评估标准定位的时间步对智能体性能的影响，本文分别采用随机投毒记为Random、Lin等[23]提出的动作偏好值记为C值、Kos等[24]提出的动作价值记为Q值及本文提出的基于TD误差的关键时间步定位方法记为TD确定关键时间步。为对比不同方法定位的时间步对学习过程的影响，对基于上述几种方法定位的关键时间步中的状态−动作对添加了大小在0.05范围内的扰动。图4展示了在Walker2D任务中，攻击不同方法定位的时间步得到的CQL中毒智能体的性能下降比例。图4中累积奖励为在测试环境中执行50次轨迹后取均值。结果表明，基于TD误差的关键时间步定位方法攻击效果最好，说明该方法能够选择出对学习过程影响更大的时间步。这是由于高价值时间步往往对应于数据集中的普遍任务模式（如自动驾驶任务中具有高价值动作为快速直行），没有直接针对学习过程中那些最具挑战性的决定性时刻，所以难以精确捕捉影响智能体长期表现的关键时间步。而本文方法关注那些模型预测性能最差的时间步，这些时刻通常是智能体在理解环境动态方面遇到的较大挑战，因此在这些关键时间步上引入扰动能够有效破坏智能体对环境的学习和适应能力，进一步揭示了智能体对这些时间步的敏感性。

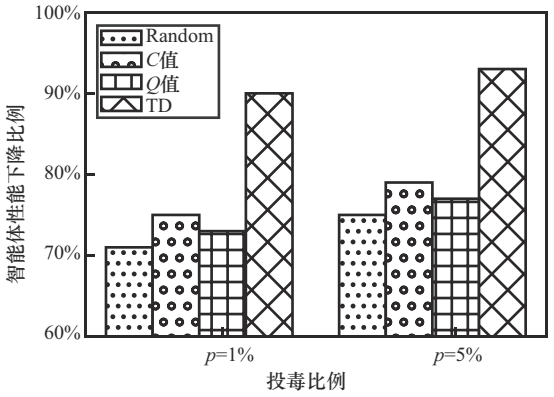

图4　不同关键点定位方法实验结果对比

2) 扰动生成方法对比。为了讨论不同投毒方法对攻击的影响，本文对Walker2D环境中训练的CQL算法智能体执行了基于不同投毒方法的攻击，

中毒智能体的性能下降比例如图5所示。关键时间步定位采用基于TD误差的关键时间步定位方法。将本文提出的基于双目标优化的投毒攻击方法（标记为DOPA）与Foley等[35]在在线场景中提出的约束范围内添加扰动的方法（标记为Perturb），扰动幅度ε上限取0.05，同时将扰动幅度固定为本文方法求出所有扰动的均值0.037（标记为Avg_Perturb）、固定为上限0.05（标记为Max_Perturb）作为对比，投毒比例p分别为1%和5%。

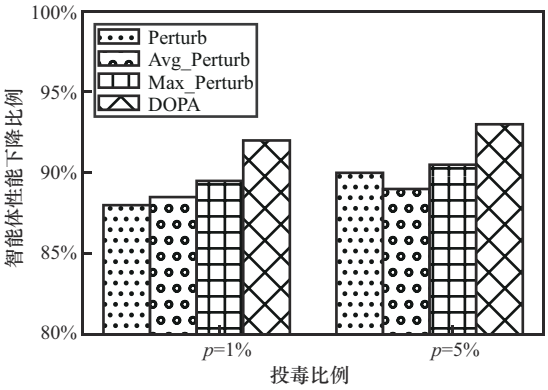

图5　不同投毒方法实验结果对比

图5中的结果表明，与在约束范围内随机添加扰动和固定扰动幅度的方法相比，本文方法在不同投毒比例下均更有效，表明优化得到的扰动能够有针对性地对智能体性能产生负面影响。基于优化的投毒过程明确了扰动生成的目标，利用模型学习过程中的关键信息，集中攻击模型的薄弱环节，放大扰动的负面影响。

另一个明显趋势为扰动幅度固定为0.05时相比均值攻击效果更好，这与表3中的结果一致，证明扰动幅度ε上限会限制攻击的性能。

3.7　不同投毒比例实验及结果分析

为了评估投毒比例p对攻击效果的影响，本文对在Walker2D环境中训练的4种算法智能体执行了不同投毒比例的攻击。图6显示了智能体性能下降比例的变化情况。投毒比例$p = 0$表示使用干净数据对智能体进行训练，且扰动幅度ε上限设置为0.05。一个明显的趋势是，随着投毒比例的增加，智能体在攻击下的性能下降比例逐渐增加。这是由于中毒数据越多，被扰动的关键时间步数量越多，从而减少了能对策略学习产生积极影响的数据数量，导致智能体性能的进一步下降。

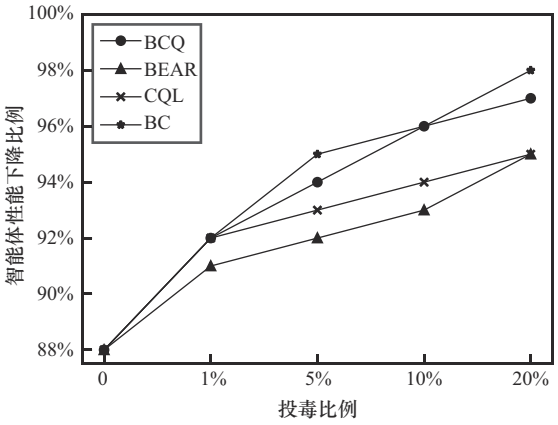

图6　不同投毒比例实验结果对比

4　结束语

本文研究了无人系统中面向离线强化学习的数据投毒攻击问题，提出了一种关键时间步动态投毒攻击方法。基于对离线强化学习过程的分析，本文确定了TD误差较大的时间步为学习过程中的薄弱环节，并将其作为投毒目标。进一步提出了基于双目标优化的投毒攻击方法，将中毒扰动的生成转化为带约束的双目标优化问题，最小化扰动幅度的同时最大化TD误差。实验结果表明，本文提出的关键时间步动态投毒攻击方法在较低代价时就能使得智能体性能大幅度下降。即使在对少部分数据集进行投毒攻击时，该方法依然表现出显著的攻击效果，展示了其在有效性和隐蔽性上的优势。

本文揭示了离线强化学习在面对数据投毒攻击时的脆弱性，强调了数据质量和安全的重要性。未来研究可以扩展本文方法，探索更多攻击场景，并开发有效的防御策略，以提升无人系统在高风险决策任务中的安全性和可靠性。

参考文献：

[1] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

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

[3] 苏新, 孟蕾蕾, 周一青, 等. 基于深度强化学习的海洋移动边缘计算卸载方法[J]. 通信学报, 2022, 43(10): 133-145.
SU X, MENG L L, ZHOU Y Q, et al. Maritime mobile edge computing offloading method based on deep reinforcement learning[J]. Journal on Communications, 2022, 43(10): 133-145.

[4] ARULKUMARAN K, DEISENROTH M P, BRUNDAGE M, et al. Deep reinforcement learning: a brief survey[J]. IEEE Signal Processing

Magazine, 2017, 34(6): 26-38.
[5] LEVINE S, KUMAR A, TUCKER G, et al. Offline reinforcement learning: tutorial, review, and perspectives on open problems[J]. arXiv Preprint, arXiv: 2005.01643, 2020.
[6] DIEHL C, SIEVERNICH T S, KRÜGER M, et al. Uncertainty-aware model-based offline reinforcement learning for automated driving[J]. IEEE Robotics and Automation Letters, 2023, 8(2): 1167-1174.
[7] CHEBOTAR Y, HAUSMAN K, LU Y, et al. Actionable models: unsupervised offline reinforcement learning of robotic skills[J]. arXiv Preprint, arXiv: 2104.07749, 2021.
[8] 乌兰, 刘全, 黄志刚, 等. 离线强化学习研究综述[J]. 计算机学报, 2024: 1-35.
WU L, LIU Q, HUANG Z G, et al. A survey of offline reinforcement learning research[J]. Chinese Journal of Computers, 2024: 1-35.
[9] SHI C, XIONG W, SHEN C, et al. Provably efficient offline reinforcement learning with perturbed data sources[J]. arXiv Preprint, arXiv: 2306.08364, 2023.
[10] KONYUSHKOVA K, ZOLNA K, AYTAR Y, et al. Semi-supervised reward learning for offline reinforcement learning[J]. arXiv Preprint, arXiv: 2012.06899, 2020.
[11] WU F, LI L Y, XU C J, et al. COPA: certifying robust policies for offline reinforcement learning against poisoning attacks[J]. arXiv Preprint, arXiv: 2203.08398, 2022.
[12] GOLDBLUM M, TSIPRAS D, XIE C L, et al. Dataset security for machine learning: data poisoning, backdoor attacks, and defenses[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 1563-1580.
[13] WU Y, MCMAHAN J, ZHU X J, et al. Reward poisoning attacks on offline multi-agent reinforcement learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(9): 10426-10434.
[14] 刘艾杉, 郭骏, 李思民, 等. 面向深度强化学习的对抗攻防综述[J]. 计算机学报, 2023, 46(8): 1553-1576.
LIU A S, GUO J, LI S M, et al. A survey on adversarial attacks and defenses for deep reinforcement learning[J]. Chinese Journal of Computers, 2023, 46(8): 1553-1576.
[15] ZHANG H, CHEN H, XIAO C, et al. Robust deep reinforcement learning against adversarial perturbations on state observations[J]. Advances in Neural Information Processing Systems, 2020, 33: 21024-21037.
[16] YANG C L, KORTYLEWSKI A, XIE C H, et al. PatchAttack: a black-box texture-based attack with reinforcement learnings[J]. arXiv Preprint, arXiv: 2004.05682, 2020.
[17] ZHANG H, CHEN H G, BONING D, et al. Robust reinforcement learning on state observations with learned optimal adversary[J]. arXiv Preprint, arXiv: 2101.08452, 2021.
[18] SUN Y, ZHENG R, LIANG Y, et al. Who is the strongest enemy? towards optimal and efficient evasion attacks in deep RL[J]. arXiv Preprint, arXiv: 2106.05087, 2021.
[19] STANDEN M, KIM J, SZABO C. SoK: adversarial machine learning attacks and defences in multi-agent reinforcement learning[J]. arXiv Preprint, arXiv: 2301.04299, 2023.
[20] ZHANG X Z, MA Y Z, SINGLA A, et al. Adaptive reward-poisoning attacks against reinforcement learning[J]. arXiv Preprint, arXiv: 2003.12613, 2020.
[21] SUN Y, HUO D, HUANG F. Vulnerability-aware poisoning mechanism for online RL with unknown dynamics[J]. arXiv Preprint, arXiv: 2009.00774, 2020.
[22] FENG J, CAI Q Z, ZHOU Z H. Learning to confuse: generating training time adversarial data with auto-encoder[J]. arXiv Preprint, arXiv: 1905.09027, 2019.
[23] LIN Y C, HONG Z W, LIAO Y H, et al. Tactics of adversarial attack on deep reinforcement learning agents[C]//Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2017: 3756-3762.
[24] KOS J, SONG D. Delving into adversarial attacks on deep policies[J]. arXiv Preprint, arXiv: 1705.06452, 2017.
[25] QU X H, SUN Z, ONG Y S, et al. Minimalistic attacks: how little it takes to fool deep reinforcement learning policies[J]. IEEE Transactions on Cognitive and Developmental Systems, 2021, 13(4): 806-817.
[26] SUN J W, ZHANG T W, XIE X F, et al. Stealthy and efficient adversarial attacks against deep reinforcement learning[J]. AAAI-20 Technical Tracks 4, 2020, 34(4): 5883-5891.
[27] YU C M, CHEN M H, LIN H T. Learning key steps to attack deep reinforcement learning agents[J]. Machine Learning, 2023, 112(5): 1499-1522.
[28] MA Y Z, ZHANG X Z, SUN W, et al. Policy poisoning in batch reinforcement learning and control[J]. arXiv Preprint, arXiv: 1910.05821 , 2019.
[29] RAKHSHA A, RADANOVIC G, DEVIDZE R, et al. Policy teaching via environment poisoning: training-time adversarial attacks against reinforcement learning[J]. arXiv Preprint, arXiv: 2003.12909, 2020.
[30] GONG C, YANG Z, BAI Y P, et al. Baffle: hiding backdoors in offline reinforcement learning datasets[C]//Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2024: 2086-2104.
[31] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. 2nd ed. Cambridge: MIT Press, 2018.
[32] WATKINS C J C H, DAYAN P. Q-learning[J]. Machine Learning, 1992, 8(3): 279-292.
[33] SCHAUL T. QUAN J, ANTONOGLOU I,et al. Prioritized experience replay[J]. arXiv Preprint, arXiv: 1511.05952, 2015.
[34] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv Preprint, arXiv: 1312.6199, 2013.
[35] FOLEY H, FOWL L H, GOLDSTEIN T, ET AL. EXECUTE ORDER 66: TARGETED DATA POISONING FOR REINFORCEMENT LEARNING[J]. arXiv Preprint, arXiv: 2201.00762, 2022.
[36] SHIN D, DRAGAN A, BROWN D S. Benchmarks and algorithms for offline preference-based reward learning[J]. arXiv Preprint, arXiv: 2301.01392, 2023.
[37] LI A, MISRA D, KOLOBOV A, et al. Survival instinct in offline reinforcement learning[J]. arXiv Preprint, arXiv: 2306.03286, 2023.
[38] KRAFT D. Algorithm 733: TOMP – Fortran modules for optimal control calculations[J]. ACM Transactions on Mathematical Software, 1994, 20(3): 262-281.
[39] FU J, KUMAR A, NACHUM O, et al. D4RL: datasets for deep data-driven reinforcement learning[J]. arXiv Preprint, arXiv: 2004.07219, 2020.
[40] TODOROV E, EREZ T, TASSA Y. MuJoCo: a physics engine for model-based control[C]//Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE Press, 2012: 5026-5033.
[41] DOSOVITSKIY A, ROS G, CODEVILLA F, et al. CARLA: an open urban driving simulator[J]. arXiv Preprint, arXiv: 1711.03938, 2017.
[42] FUJIMOTO S, MEGER D, PRECUP D. Off-policy deep reinforcement learning without exploration[J]. arXiv Preprint, arXiv: 1812.02900, 2018.
[43] KUMAR A, FU J, TUCKER, G, et al. Stabilizing off-policy q-learning

via bootstrapping error reduction[J]. arXiv Preprint, arXiv: 1906.00949, 2019.
[44] KUMAR A, ZHOU A, TUCKER G, et al. Conservative q-learning for offline reinforcement learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 1179-1191.
[45] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv Preprint, arXiv: 1707.06347, 2017.

[作者简介]

周雪（1994– ），女，黑龙江牡丹江人，哈尔滨工程大学博士生，主要研究方向为网络安全、人工智能安全、强化学习。

苘大鹏（1980– ），男，辽宁抚顺人，博士，哈尔滨工程大学教授、博士生导师，主要研究方向为网络流量安全监测、新型网络与人工智能安全。

许晨（1996– ），男，山东菏泽人，博士，哈尔滨工程大学讲师、硕士生导师，主要研究方向为人工智能安全、自然语言处理、语音处理。

吕继光（1987– ），男，黑龙江哈尔滨人，哈尔滨工程大学副教授、硕士生导师，主要研究方向为工业互联网应用安全、人工智能安全。

曾凡一（1993– ），女，蒙古族，辽宁昌图人，哈尔滨工程大学博士生，主要研究方向为网络入侵检测、加密恶意流量分析。

高朝阳（1999– ），男，河南驻马店人，哈尔滨工程大学硕士生，主要研究方向为网络安全、人工智能安全。

杨武（1974– ），男，辽宁宽甸人，博士，哈尔滨工程大学教授、博士生导师，主要研究方向为网络与信息安全、人工智能应用及安全。