智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于自适应分位数的离线强化学习算法

周娴玮, 王宇翔, 罗仕鑫, 余松森

引用本文:
周娴玮, 王宇翔, 罗仕鑫, 等. 基于自适应分位数的离线强化学习算法[J]. 智能系统学报, 2025, 20(5): 1093-1102.
ZHOU Xianwei, WANG Yuxiang, LUO Shixin, et al. Offline reinforcement learning with adaptive quantile[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(5): 1093-1102.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202410016

DOI: 10.11992/tis.202410016
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20250623.1537.006

基于自适应分位数的离线强化学习算法

周娴玮, 王宇翔, 罗仕鑫, 余松森
(华南师范大学 人工智能学院, 广东 佛山 528225)

摘　要: 离线强化学习旨在仅通过使用预先收集的离线数据集进行策略的有效学习, 从而减少与环境直接交互所带来的高昂成本。然而, 由于缺少环境对智能体行为的交互反馈, 从离线数据集中学习到的策略可能会遇到数据分布偏移的问题, 进而导致外推误差的不断加剧。当前方法多采用策略约束或模仿学习方法来缓解这一问题, 但其学习到的策略通常较为保守。针对上述难题, 提出一种基于自适应分位数的方法。具体而言, 该方法在双 Q 估计的基础上进一步利用双 Q 的估计差值大小对分布外未知动作的价值高估情况进行评估, 同时结合分位数思想自适应调整分位数来校正过估计偏差。此外, 构建分位数优势函数作为策略约束项权重以平衡智能体对数据集的探索和模仿, 从而缓解策略学习的保守性。最后在 D4RL (datasets for deep data-driven reinforcement learning) 数据集上验证算法的有效性, 该算法在多个任务数据集上表现优异, 同时展现出在不同场景应用下的广泛潜力。

关键词: 离线强化学习; 分布偏移; 外推误差; 策略约束; 模仿学习; 双 Q 估计; 价值高估; 分位数
中图分类号: TP301.6　　**文献标志码:** A　　**文章编号:** 1673−4785(2025)05−1093−10

中文引用格式: 周娴玮, 王宇翔, 罗仕鑫, 等. 基于自适应分位数的离线强化学习算法 [J]. 智能系统学报, 2025, 20(5): 1093–1102.
英文引用格式: ZHOU Xianwei, WANG Yuxiang, LUO Shixin, et al. Offline reinforcement learning with adaptive quantile[J]. CAAI transactions on intelligent systems, 2025, 20(5): 1093–1102.

Offline reinforcement learning with adaptive quantile

ZHOU Xianwei, WANG Yuxiang, LUO Shixin, YU Songsen
(School of Artificial Intelligence, South China Normal University, Foshan 528225, China)

Abstract: Offline reinforcement learning aims to reduce the high cost of environmental interaction by learning effective policies solely from precollected offline datasets. However, the absence of interactive feedback can cause a distribution shift between the learned policy and the offline dataset, leading to increased extrapolation errors. Most existing methods address this problem using policy constraints or imitation learning, but they often result in overly conservative policies. To address the above problems, an adaptive quantile-based method is proposed. Building upon dual Q-estimation, the relationship between dual Q-estimates is further analyzed, using their differences to assess overestimation in out-of-distribution actions. The quantile is then adaptively adjusted to correct bias overestimation. Additionally, a quantile advantage is introduced, which serves as a weight for the policy constraint term, balancing exploration and imitation to reduce policy conservativeness. Finally, the proposed approach is validated on the D4RL dataset, where it achieves excellent performance across multiple tasks, showing its potential for broad application in various scenarios.

Keywords: offline reinforcement learning; distribution shift; extrapolation error; policy constraint; imitation learning; double Q-estimation; overestimation; quantile

随着人工智能技术的发展, 深度强化学习技术展现出广泛的应用潜力, 如在教育[1]、医疗[2]、游戏博弈[3-4]、路径规划[5-6] 和机器人控制领域[7-8]等。现实应用场景中, 与环境进行直接交互通常是代价高昂的, 而离线强化学习的适时出现为应对这一难题提供了思路。离线强化学习专注于从预先收集的离线数据集中学习最优策略, 无需在学习过程中与环境进行交互。这种方法突破了传统在线强化学习的限制, 允许智能体在没有即时环境反馈的情况下, 通过学习历史数据来提高决策能力[9]。在自动驾驶等领域中, 与环境的交互往往代价高昂或风险巨大。离线强化学习通过利用已有的离线数据, 提高策略学习的速度和安全性, 为解决这一问题提供了有效方案[10]。

然而, 离线强化学习也面临着独特的挑战。由于智能体缺少与环境的交互, 学习策略往往与

收稿日期: 2024−10−12.　**网络出版日期:** 2025−06−23.
基金项目: 广东省应用型科技研发重大专项 (2016B020244003); 广东省企业科技特派员项目 (GDKTP2020014000); 广东省基础与应用基础研究基金项目 (2020B1515120089, 2020A1515110783).
通信作者: 余松森. E-mail: yss8109@163.com.

离线数据集中的数据产生分布偏移，进而导致算法在分布外的未知状态中做出错误的决策，引起外推误差的加剧[11-12]。为应对这一挑战，现有的工作大多采用策略约束思想，通过采用不同的距离度量，如 KL 散度 (Kullback-Leibler divergence)[13]、MMD(maximum mean discrepancy)[14] 作为约束项，显式或隐式地将学习策略限制到离线数据集中已知的行为策略上来缓解分布偏移问题。另一种解决方案是采用模仿学习的方法，对离线数据集中的劣质数据进行过滤，进而模仿数据集中的更优动作来寻求最优策略，同时也避免了陷入分布外未知状态的问题[15]。但是，离线强化学习通常对离线数据集有较强的依赖性。若离线数据集质量不佳，过强的约束或模仿往往致使学习策略过于保守。

本文专注于解决分布偏移导致的外推误差问题以及平衡智能体对离线数据集的探索和模仿。现有算法普遍使用双 Q 估计中的较小值作为动作价值的估计[16]，从而在一定程度上缓解价值高估问题。然而，当智能体采取离线数据集分布之外的未知动作时，采用双 Q 估计的较小值仍会出现动作价值的异常高估问题，且单个估计值受限于数据集的轨迹质量。为此，本文考虑到进一步结合双 Q 的估计差值大小来评估动作价值的高估情况。当双 Q 的估计差值较大时，其在一定程度上反映智能体大概率陷入分布外的未知状态，因而展现出对动作价值的异常估计。

结合以上所述思想，本文提出一种基于自适应分位数的离线强化学习算法。依据双 Q 估计的绝对误差大小变化来评估训练中动作价值的高估情况，并自适应调整分位数大小，从而对异常估计偏差进行校准。同时，为缓解策略学习的保守性，将自适应分位数结合动作优势来构建分位数优势，以此对策略约束项加权，从而实现智能体对离线数据集探索和模仿的平衡。最后，在 D4RL (datasets for deep data-driven reinforcement learning)[17] 基准上对算法进行测试，实验结果表明该算法展现出了先进的性能表现和良好的普适性。综上所述，本文的主要贡献体现在如下方面：1) 在双 Q 估计的基础上进一步提出利用双 Q 估计的估计差值大小来反映动作价值的高估情况，并据此自适应调节分位数来校准偏差，从而缓解外推误差的加剧，提高算法的泛化性。2) 结合现有的策略约束方法，创新地提出了使用分位数优势对约束项进行加权，旨在平衡智能体对离线数据集的探索和模仿，进而缓解策略学习的保守性。3) 在 D4RL 数据集上进行了算法验证和实验对比，该算法取得了良好的性能表现，尤其在随机环境和中等环境下表现较优，同时展现了在多样化环境中的适用性，为后续的方法研究提供了一定参考。

1　现状分析

离线强化学习作为一种不依赖实时环境交互的强化学习方法，在降低成本和风险方面具有显著优势，适用于数据收集成本高昂或存在风险的场景。外推误差问题是离线强化学习面临的重要挑战，即在决策过程中，智能体可能采用数据集之外的未知动作，产生价值的过高估计，从而导致性能下降。策略约束和模仿学习是当前应对这一难题的两类主要方法[18]。

策略约束方法通过显式或隐式地将智能体的学习策略约束到数据集的行为策略上，来保持策略的更新始终处于安全区域内。显式策略约束通过直接估计数据集的行为策略，并将无约束的策略优化目标修改为带约束的策略优化目标来解决分布偏移问题，同时引入不同的指标来度量两种数据分布的距离。隐式策略约束不依赖对数据集行为策略的估计，而是使用数据集中的样本改进优化目标来隐式地限制学习策略。BRAC(behavior regularized actor critic)[13] 方法提出在策略评估过程或策略改进步骤中，从优化目标中减去一个分歧项作为对分布外未知动作的惩罚，并在实验中使用 KL 散度作为衡量数据分布间的距离。为了简化日益复杂的离线强化算法，TD3+BC[19] 算法提出，在 TD3(twin delayed deep deterministic policy gradient)[16] 算法中的策略改进目标上添加一个行为克隆项 (behavior cloning, BC)，其在 D4RL 数据集上展现出了先进的性能，且避免了算法的复杂性。wPC(weighted policy constraints)[20] 算法之后在 TD3+BC 算法的基础上进行了改进，提出对策略约束项进行自适应加权，消除对学习策略的不利约束，同时保留对理想行为的必要约束，在不引入额外超参数的基础上实现了性能的提高。

模仿学习的核心在于模仿数据集中已知的行为策略。最简单的模仿学习方法是行为克隆，可实现对数据集中动作的精确复制。由于离线数据集中数据的多样性，当前的模仿学习方法倾向于先过滤掉数据集中的劣质动作，然后模仿较优行为[21-23]。CRR(critic regularized regression)[15] 算法通过设置保守的优势估计器来评估离线数据集中动作的优劣性，并使用指示函数来过滤掉低于平均水平的劣质动作，从而实现对数据集中优秀动作的模仿。QFIL(quantile filtered imitation learning)[24] 算法提出一种可以与任何值函数估计技术相结合

的策略改进算子, 通过定义值函数分位数来实现分位数过滤模仿学习, 从而模仿数据集上表现优秀的动作。同时, QFIL 还指出在任何使用优势函数的情况下, 都可以使用值函数分位数优势来代替标准值函数优势。

现有方法从不同的角度对外推误差问题进行解决, 但其学习到的策略通常具有一定程度的保守性[20]。为了平衡缓解外推误差和策略保守性问题, 本文提出一种全新的方法, 即基于双 Q 估计的估计差值来评估价值的高估情况, 进而自适应调节分位数校准偏差。同时定义分位数优势加权来避免对离线数据集中的劣质动作进行模仿, 缓解学习策略的保守性, 从而提高离线强化学习算法的性能表现。

2　预备知识

2.1　离线强化学习

离线强化学习旨在利用预先收集的历史数据进行策略学习, 而无需与环境进行直接交互。如图 1 所示, 在离线强化学习中, 智能体仅从存储离线数据的数据缓冲区进行最优策略的学习, 且在学习期间无法与环境进行直接的交互。

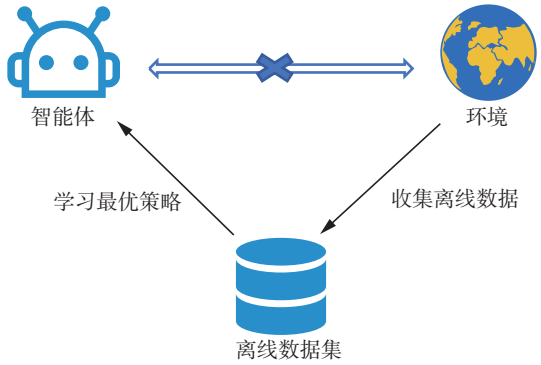


图 1　离线强化学习示意

Fig. 1　Illustration of offline reinforcement learning

尽管离线强化学习不依赖于实时交互, 但其与在线强化学习有着相同的核心目标: 在给定的环境任务中求解出最优策略, 同时最大化累积奖励, 即

$$\pi_\theta \leftarrow \arg\max_{\pi_\theta} E_{s\sim D}[Q(s,\pi_\theta(s))]$$

式中: π_θ是当前状态 s 下学习到的策略, Q 是对采取当前策略的动作价值估计, D 是离线数据集。然而由于离线强化学习无法通过交互获得环境的信息反馈, 导致其存在“外推误差”问题, 即在策略学习过程中, 智能体会采取离线数据集之外的未知动作, 通过不断地学习迭代偏离已知的数据分布, 同时产生对未知动作的价值过估计, 最终导致性能的下降。为缓解这一问题, 现有方法提出将学习策略约束到离线数据集的行为策略上, 通过在原本策略优化目标函数的基础上增加一个额外的惩罚项来实现[18], 即

$$\pi_\theta \leftarrow \arg\max_{\pi_\theta} E_{(s,a)\sim D}[Q(s,\pi_\theta(s)) - \lambda \cdot d(\pi_\theta, a)]$$

式中: d 表示两种策略之间距离的度量, λ用来控制约束项的强度, (s, a) 是离线数据集中的状态动作对。$Q(s,\pi_\theta(s))$旨在进行学习策略的优化, $-\lambda \cdot d(\pi_\theta, a)$则保证策略在安全区域内进行改进更新, 这有效地减少了离线强化学习中外推误差的进一步传递。

2.2　分位数回归

分位数是连续分布函数中的一个点, 这个点对应概率 q, 也叫作分位点。分位数表示数据百分比低于该值的值。如图 2 所示, x_0 是随机变量 X 的一个值, 存在概率 $P(X < x_0)=q$, 则 x_0 是 X 的 q 分位数。当 q =0.5 时, 那么 x_0 就是中位数, 表示随机变量 X 有 50% 的概率小于或等于这个值, 50% 的概率大于这个值。

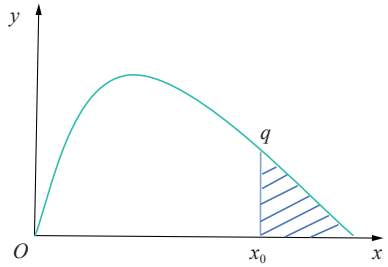


图 2　分位数介绍

Fig. 2　Illustration of quantile

分位数回归[25] 是一种广泛应用于经济学领域的回归分析方法。与传统的最小二乘回归不同, 分位数回归是基于绝对偏差的优化, 更关注数据的分布情况, 对异常值有较强的鲁棒性。对于 $q \in (0,1)$, 其分位数回归损失是一种非对称的损失函数, 具体为

$$\text{loss}(y,\hat{y}) = (1-q)\cdot\max(0,\hat{y}-y) + q\cdot\max(0,y-\hat{y})$$

式中: y 表示真实值, $\hat{y}$表示预测值; 用权重 q 惩罚低估误差, 权重 $1-q$ 惩罚高估误差, 以此来平衡预测值的低估和高估。通过不同分位权重的损失函数, 模型训练可以针对特定类型的误差进行优化, 从而提高模型的整体性能。同时, 在数据分布不均匀或存在异常值的情况下, 适当调整分位数可以使模型更加稳健, 减少异常值对模型性能的影响。

3　本文方法

本章将详细介绍所提出的基于自适应分位数

的方法。由于在离线强化学习中智能体会出现对未知动作的不准确估计问题，现有方法普遍采用双 Q 或多 Q 估计[26]来缓解价值的高估。双 Q 估计是 TD3 算法[16]中提出的一项关键技术，旨在缓解在线强化学习中的价值高估问题。其核心思想是采用两个相同网络架构的价值网络，在计算目标值时取二者中的较小值，即

$$y = r(s,a) + \gamma \times \min\left(Q_{\phi_1}(s',a'), Q_{\phi_2}(s',a')\right)$$

式中：$r(s,a)$ 表示在状态 s 下执行动作 a 获得的即时奖励，γ 表示折扣因子，$Q_{\phi_1}(s',a')$、$Q_{\phi_2}(s',a')$表示价值的双 Q 估计。这种方法有效地缓解了过估计偏差，在在线强化学习中的许多连续控制任务上都取得了较优的表现。

然而在离线强化学习中，智能体不能与环境进行直接交互，进而无法得到环境的信息反馈，这也限制了双 Q 估计在解决过估计问题上的表现。考虑到由于在双 Q 估计中仅额外增加了一个相同网络架构的价值网络，而两个价值网络的输出并无直接关联，同时仅使用双 Q 的较小值也造成另一网络输出值的空闲。故本文旨在结合双 Q 估计的输出值关系设计合理的指标以进一步缓解价值过估计问题。

本文考虑到将双 Q 估计的估计差值大小作为衡量价值过估计问题的度量。智能体在学习中会出现如下现象：如图 3(a) 所示，$Q_{\min}$ 和 $Q_{\max}$ 表示在对动作进行双 Q 估计时的较小估计值和较大估计值。随着学习的迭代进行，Q 值的估计会不断递增，同时双 Q 的估计差值 $Q_{\max}-Q_{\min}$ 也会变大。图 3(b) 反映了智能体采取相应学习策略后的平均回报值变化。在策略学习初期，智能体会倾向于选择数据集中动作价值更高的行为，从而相应学习策略的回报会不断增加，直至达到最高点 P。然而，随着学习回合的增加，图 3(a) 中价值估计的偏差会越来越大，此时对于同一动作的两个 Q 值之间估计差值却异常大，这一定程度上反映出产生了动作价值的异常估计问题，导致智能体选择数据集之外的未知动作，从而引起图 3(b) 中 P 点之后策略性能的不断下降。

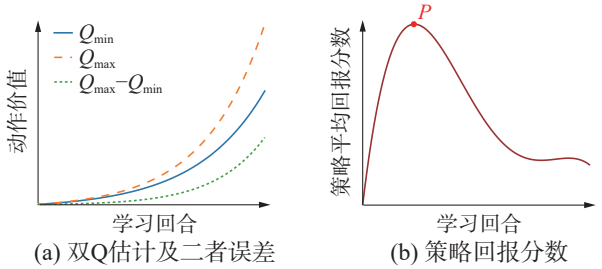


图 3　双 Q 误差及策略回报变化

Fig. 3　Curve of double-Q error and policy return

针对这一问题，本文利用双 Q 估计差值大小的变化情况，并将其结合分位数思想，提出基于自适应分位数的离线强化学习方法，通过使用双 Q 的估计差值自适应调节分位数来平衡对价值估计的高估和低估惩罚。此外，为缓解学习策略的保守性，在自适应分位数的基础上引入分位数优势加权来平衡智能体对离线数据集的探索和模仿。本文的算法框架如图 4 所示。

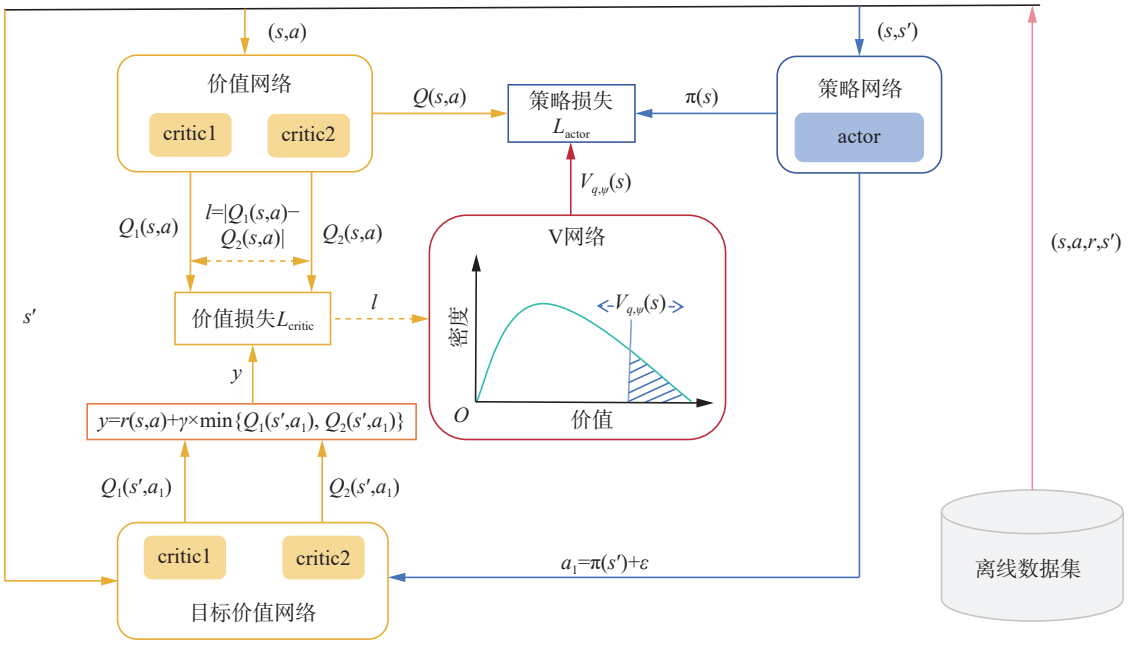


图 4　算法框架

Fig. 4　Algorithm framework

离线强化学习算法的关键在于价值网络 critic 和策略网络 actor 的优化更新。如图 4 所示, 本文的核心包括结合双 Q 的估计差值 l 来优化 V 网络和将值函数分位数应用到策略优化中这两部分。

3.1　自适应分位数

在离线强化学习中, 分位数主要用来调节策略以适应数据分布的不同情况, 减少外推误差, 提高策略的鲁棒性。现有工作多数使用预先设定好的固定分位数[24], 对人工调参经验的要求较高, 且对离线数据集有极大依赖性。同时, 使用固定的分位数, 如中位数 0.5, 四分位数 0.25 和 0.75 等, 可能无法充分捕捉数据的复杂性或满足特定的分析需求。为此, 本文提出基于双 Q 估计差值进行分位数自适应调节, 动态地选择合适的分位数, 在学习过程中校准估计偏差, 以优化算法性能。同时鉴于分位数 q 的取值范围介于 0、1 之间, 对区间进行归一化处理。本文设计分位数 q 计算公式为

$$q = e^{-\beta\left|Q_{\phi_1}(s,a)-Q_{\phi_2}(s,a)\right|}$$

式中: $\beta>0$, 是一个超参数, 用来控制双 Q 的估计差值的影响强度; $Q_{\phi_1}(s,a)$和$Q_{\phi_2}(s,a)$是双 Q 网络对状态动作对 (s,a) 的动作价值估计。当双 Q 估计的估计差值过高时, 智能体可能陷入分布外未知动作, 进而对动作价值产生过估计问题。为校准这种高估偏差, 将分位数自适应减小, 从而加强对高估误差的惩罚。相反地, 当双 Q 估计差值较小时, 说明其对动作价值的估计是较为准确的, 此时希望智能体采取更高价值的动作, 将分位数自适应增加, 加强对保守低估的惩罚。总的来说, 自适应分位数可以在智能体寻求更优策略的同时, 减少来自未知动作的过估计影响, 从而缓解学习中的外推误差问题, 同时提高算法在不同数据场景下的泛化能力。

3.2　分位数优势加权

为缓解策略约束所导致的学习策略保守性, 结合上一小节的自适应分位数构建分位数优势函数作为策略约束权重, 从而避免对离线数据集中的劣质动作进行模仿。在离线强化学习中, 分位数通常与状态 s 的价值期望相结合得到价值函数分位数[24]。为此, 引入一个 V 网络来拟合价值函数分位数, 并将自适应分位数思想结合到网络参数优化中。同时为避免分位数回归损失中绝对值损失所导致的梯度问题, 目标函数优化使用非对称的二次损失函数, 其在现有的 IQL(implicit q-learning)[27] 方法中也有所应用。最后, 设计优化 V 网络的目标损失函数:

$$L_{\mathrm{Vnet}}(\psi) = E_{(s,a)\sim D}\left[L_2^q\left(Q_{\phi'}(s,a)-V_{q,\psi}(s)\right)\right]$$

式中: 对于函数$L_2^q(\mu)$, 当$\mu\geqslant 0$时, 其取值为$q\mu^2$; 当$\mu<0$时, 其取值为$(1-q)\mu^2$。通过这种非对称损失结合自适应分位数来实现对价值估计偏差的动态校正。之后, 结合 wPC 方法中提出的加权策略约束思想, 使用分位数优势对约束进行加权, 与 TD3 算法相结合, 最终构造优化策略的目标损失函数为

$$L_{\mathrm{actor}}(\theta) = E_{(s,a)\sim D}[-\lambda Q(s,\pi_\theta(s))+ \mathbb{1}\left(Q_\phi(s,a)\geqslant V_{q,\psi}(s)\right)\cdot(\pi_\theta(s)-a)^2]$$

式中: $\mathbb{1}$为指示函数, λ沿用 TD3+BC 方法中的计算方式, 即

$$\lambda = \frac{\alpha N}{\sum_{(s_i,a_i)}|Q(s_i,a_i)|}$$

式中: α是一个超参数, N是批大小。与 wPC 方法相比, 本文所提出的分位数优势加权方法改进了策略约束项的权重, 即利用基于自适应分位数的值函数分位数来代替动作价值的期望, 更易于捕捉数据分布的变化。最后, 算法具体流程见算法 1。

算法 1　基于自适应分位数的离线强化学习算法

输入　离线数据集 D, 训练回合 T, 采样批次大小 B, 折扣因子 γ, 策略噪声 ϵ, 软更新参数 τ, actor 网络 π_θ, critic 网络Q_{ϕ_1}, Q_{ϕ_2}, V 网络 $V\psi$。

输出　策略回报的归一化分数。

1) 随机初始化 actor 网络参数权重 θ, critic 网络权重参数 ϕ_1、ϕ_2, V 网络权重参数 ψ;

2) 初始化目标网络参数;

3)**While** 迭代轮次 $t\in[0,T]$ **do**

4) 从离线数据集 D 中采样批量数据 B:

$$B=(s,a,r,s')$$

5) 从策略中选取动作: $a'=\pi_\theta(s')+\epsilon$;

6) 计算目标价值:

$$y = r(s,a)+\gamma\times\min\left(Q_{\phi_1}(s',a'),Q_{\phi_2}(s',a')\right);$$

7) 更新 critic 网络参数:

$$\phi_i \leftarrow \arg\min_{\phi_i}\left(y-Q_{\phi_i}(s,a)\right)^2;$$

8) 利用式 (1) 计算分位数 q;

9) 利用式 (2) 更新 V 网络参数;

10) 利用式 (3) 更新 actor 网络参数;

11) 对目标网络进行软更新:

$$\theta' \leftarrow \tau\theta+(1-\tau)\theta'$$

$$\phi_1' \leftarrow \tau\phi_1+(1-\tau)\phi_1'$$

$$\phi_2' \leftarrow \tau\phi_2 + (1-\tau)\phi_2'$$

12) **End**

4　实验验证

4.1　实验设置

实验中主要的超参数设置如表 1 所示。实验所使用的离线数据集为 D4RL 基准中的 halfcheetah、hopper 和 walker2d 3 类运动控制任务数据集(如图 5 所示)。本文所有的实验均在“-v2”版本环境下进行评估。实验中使用到的神经网络，包括 actor 网络、critic 网络和 V 网络，其结构设计及输入输出如图 6 所示。

同时，为方便与现有离线强化学习方法进行对比，实验中使用了归一化分数 $S_{\text{normalized}}$ 来代替平均回报。归一化分数的具体计算公式为

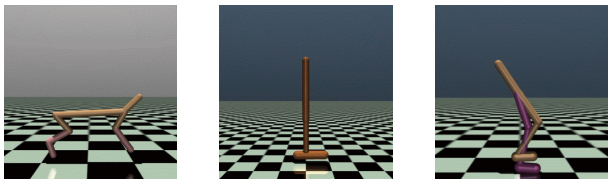
$$S_{\text{normalized}} = \frac{S_{\text{policy}} - S_{\text{random}}}{S_{\text{expert}} - S_{\text{random}}} \times 100$$

式中：S_{policy} 代表学习策略在数据集上的性能分数，S_{random} 和 S_{expert} 分别表示随机策略和专家级策略的性能表现。

表 1　超参数设置

Table 1　Hyperparameter setting

参数名称	参数值
采样批大小	256
学习率	3×10^{-4}
优化器	Adam
折扣因子γ	0.99
策略噪声ϵ	0.2
噪声截断大小	0.5
软更新系数τ	0.005
策略更新频率	2
双Q差值控制强度β	{0.3, 3}
约束控制系数α	2.5



(a) Halfcheetah　(b) Hopper　(c) Walker2d

图 5　仿真环境

Fig. 5　Simulation environment

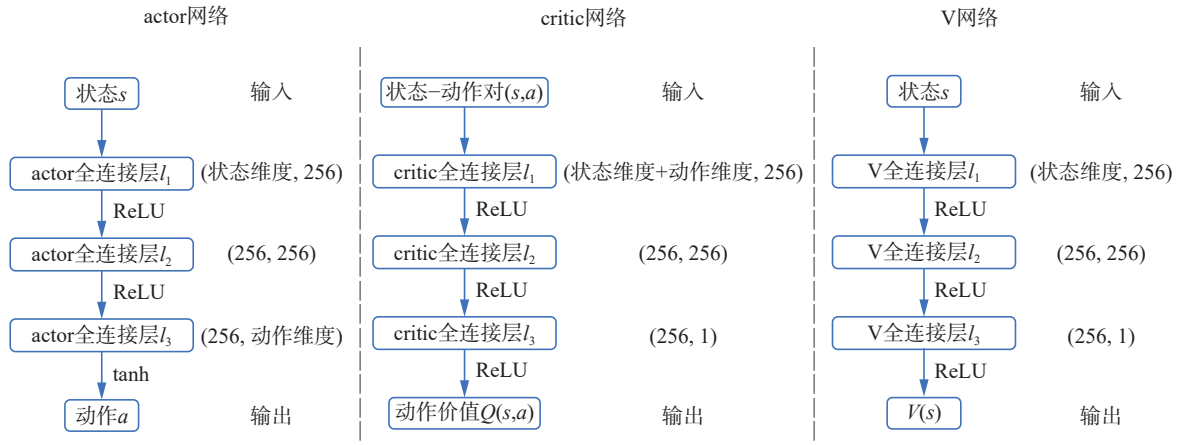


图 6　网络结构

Fig. 6　Network structure

4.2　在 D4RL 上的性能表现

D4RL 是评估离线强化学习算法的主要数据集之一，其包含了一系列广泛任务和多种丰富的数据集[17]。本文在 D4RL 数据集中 halfcheetah、hopper 和 walker2d 3 种运动控制任务环境对所提出的方法进行了实验评估和验证，并与当前的一些先进离线强化学习算法进行了对比，具体如表 2 所示。其中，‘r’、‘m’、‘m-r’和‘m-e’分别表示 random、medium、medium-replay 和 medium-expert 这 4 类不同的数据集。random 数据集使用来自随机初始化策略与环境进行交互产生的 1×10^6 个样本；medium 数据集使用经过训练的性能约为专家策略 1/3 的策略与环境交互产生的 1×10^6 个样本；medium-replay 数据集是中等策略在训练中使用经验回放缓冲区产生样本；在 medium-expert 数据集中，专家策略与环境交互产生的样本和中等策略与环境交互产生的样本各占 50%。在实验中，由于 medium-expert 数据集包含的离线数据质量较优，参数 β 的值设置为 3，在其余任务环境下设置为 0.3。

从表 2 的实验结果中可以看出，本文方法在诸多任务上表现良好，同时所提方法在 2 个 random 数据集和 3 个 medium 数据集上相对于其他离线强化学习方法都取得了较优的性能表现，且在

“halfcheetah-r”、“hopper-r”以及“hopper-m”上的性能提升幅度较为明显。对上述实验现象进行分析并归纳原因如下: 1) 本文方法基于双 Q 估计差值来评估智能体出现价值异常估计情况, 通过分位数自适应来校准偏差, 同时动态调节对高估误差的惩罚力度, 使过估计问题得到一定的缓解, 从而提高了算法在不同数据场景下的泛化能力。2) 使用值函数的分位数优势函数对策略约束进行加权, 一定程度上避免了对离线数据集中的劣质动作进行模仿, 从而缓解了学习策略的保守性问题, 因此在劣质数据较多的 random 和 medium 数据集下有着较大的性能表现提升。

表 2　不同离线强化学习算法在 D4RL 上性能表现对比

Table 2　Performance comparison of different offline reinforcement learning algorithms on D4RL

任务名(-v2)	BC[18]	TD3+BC[19]	wPC[20]	IQL[27]	CQL[28]	RvS[29]	%BC[30]	DT[30]	本文方法
halfcheetah-r	2.3	11.0	19.6	11.2	18.6	3.9	2.0	2.2	**23.9**
hopper-r	4.8	8.5	19.9	7.9	9.1	0.2	4.1	7.5	**22.5**
walker2d-r	1.7	1.6	0.7	5.9	2.5	**7.7**	1.7	2.0	1.2
halfcheetah-m	42.6	48.3	53.3	47.4	49.1	41.6	42.5	42.6	**55.3**
hopper-m	52.9	59.3	86.5	66.3	64.6	60.2	56.9	67.6	**99.8**
walker2d-m	75.3	83.7	86.0	78.3	82.9	71.7	75.0	74.0	**86.3**
halfcheetah-m-r	36.6	44.6	48.3	44.2	47.3	38.0	40.6	36.6	**48.5**
hopper-m-r	18.1	60.9	97.0	94.7	**97.8**	73.5	75.9	82.7	95.2
walker2d-m-r	26.0	81.8	**89.9**	73.9	86.1	60.6	62.5	66.6	86.2
halfcheetah-m-e	55.2	90.7	**93.7**	86.7	85.8	92.2	92.9	86.8	92.4
hopper-m-e	52.5	98.0	95.7	91.5	102.0	101.7	**110.9**	107.6	99.6
walker2d-m-e	107.5	110.1	**110.7**	109.1	109.5	106.0	109.0	108.1	110.6
总计	475.5	698.5	800.7	717.1	755.5	657.3	674.0	684.3	**821.5**

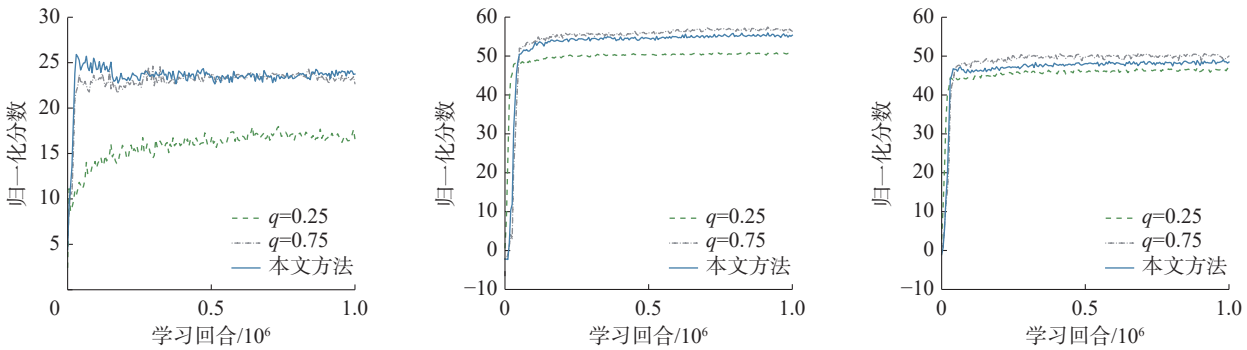
4.3　对比实验

为验证所提出的自适应分位数思想更具普适性, 现将式 (1) 中分位自适应调节修改为预先设定的固定分位数并进行对比。当前基于分位数的离线强化学习算法大多将表示分位的参数设置为 q=0.75[24]。故将所提出的自适应分位数方法与其进行对比, 并额外设置 q=0.25 进行实验对照。由于在实际应用中希望能够使用更少的专家数据来学习出最优的策略, 故本文在 random、medium 和 medium-replay 3 种数据集上进行实验对比, 实验结果见表 3, 学习曲线如图 7 所示。

表 3　不同分位数设置的性能表现对比

Table 3　Performance of different quantile settings

任务名(-v2)	q=0.25	q=0.75	本文方法
halfcheetah-r	16.8	23.1	**23.9**
hopper-r	7.8	20.5	**22.5**
walker2d-r	0.5	1.4	1.2
halfcheetah-m	50.6	**56.6**	55.3
hopper-m	70.3	96.4	**99.8**
walker2d-m	85.3	23.9	**86.3**
halfcheetah-m-r	46.4	**49.6**	48.5
hopper-m-r	88.3	**96.0**	95.2
walker2d-m-r	85.6	85.2	**86.2**
总计	451.6	452.7	**518.9**



(a) halfcheetah-r　(b) halfcheetah-m　(c) halfcheetah-m-r

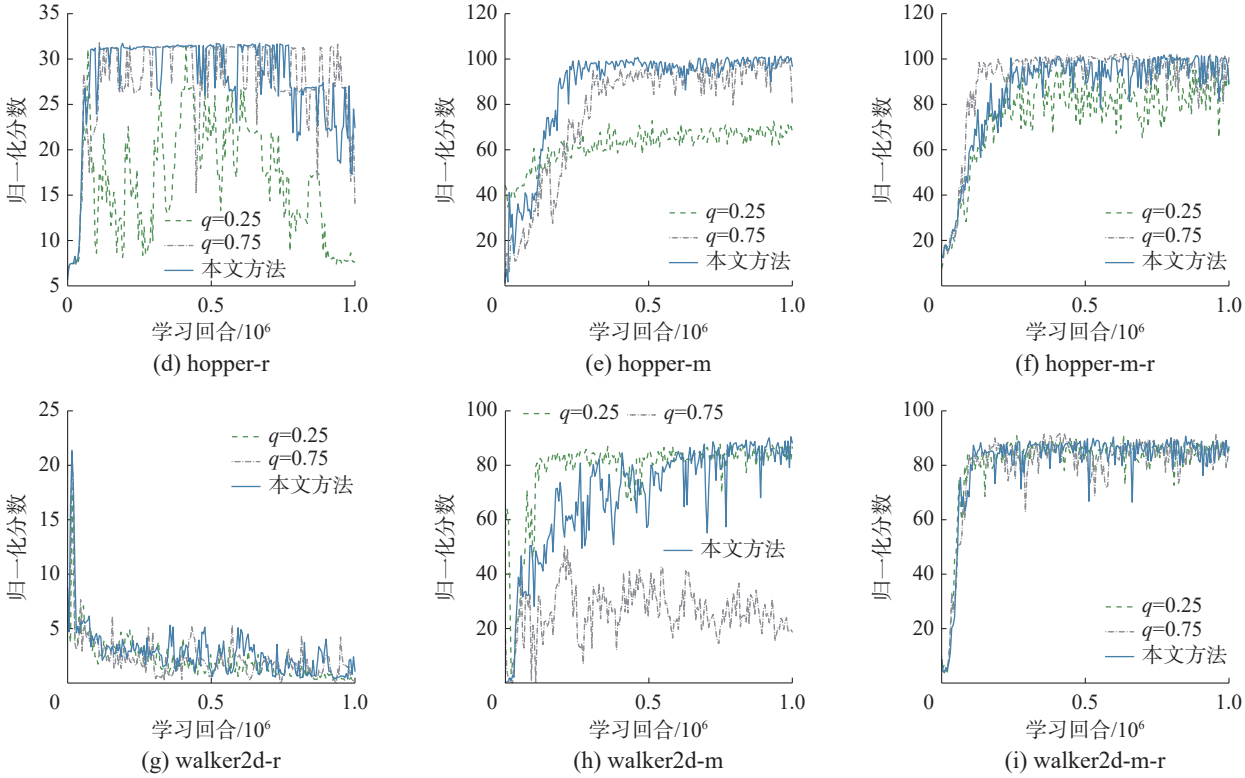


图 7　不同分位数设置下学习曲线对比

Fig. 7　Comparison of learning curves under different quantile setting

从表 3 的实验结果可以看出，所提出的自适应分位数方法表现是更优的。在 q=0.75 的分位数设置下，尽管其在“halfcheetah-r”和“hopper”表现也较为良好，但在 walker2d-m 数据集下，其得分只有 23.9。而在 q=0.25 的分位数设置下，其在各数据集上的表现都较差。同时如图 7 所示，q=0.25 的情况下，在大部分数据集上其学习曲线总是先收敛，性能表现有限。而 q=0.75 和所提方法的学习曲线在大部分数据集上都探索到了更优的策略，但 q=0.75 在“walker2d-m”数据集上表现异常。这是由于预先设定的固定分位数都是对特定的估计误差进行较强的惩罚，容易在学习中出现探索过度或模仿过度的问题，从而导致最终性能表现不佳，且直接选择合适的分位数对调参经验要求较高。而使用自适应分位数可以根据学习情况进行高估偏差和低估偏差的平衡，并相应调整惩罚力度，从而使算法在广泛任务上具有更优的普适性。

5　结束语

本文针对现有离线强化学习中存在的外推误差和学习策略保守性问题，创新地提出一种基于自适应分位数的离线强化学习算法。通过双 Q 估计的估计差值来评估动作价值的高估情况，从而自适应调节分位数，校准估计偏差来提高算法的泛化能力。同时构建分位数优势加权来缓解策略约束过强导致的保守性问题，从而避免对离线数据集中的劣质动作进行模仿。最后，通过在 D4RL 数据集上与现有先进离线强化学习方法以及固定分位数设置的实验对比，有效验证了该算法具有更优的性能表现和普适性。考虑到现实应用中对离线数据集的数据多样性要求较高，因此未来工作将聚焦于探索更优的分位数适应方式，进一步提高算法的泛化性。

参考文献：

[1] SINGLA A, RAFFERTY A N, RADANOVIC G, et al. Reinforcement learning for education: opportunities and challenges[EB/OL]. (2021−07−15)[2024−10−12]. https://arxiv.org/abs/2107.08828v1.

[2] LIU Siqi, SEE K C, NGIAM K Y, et al. Reinforcement learning for clinical decision support in critical care: comprehensive review[J]. Journal of medical Internet research, 2020, 22(7): e18477.

[3] VINYALS O, BABUSCHKIN I, CZARNECKI W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning[J]. Nature, 2019, 575(7782): 350−354.

[4] 李霞丽, 王昭琦, 刘博, 等. 麻将博弈 AI 构建方法综述

[J]. 智能系统学报, 2023, 18(6): 1143−1155.
LI Xiali, WANG Zhaoqi, LIU Bo, et al. Survey of Mahjong game AI construction methods[J]. CAAI transactions on intelligent systems, 2023, 18(6): 1143−1155.

[5] 朱少凯, 孟庆浩, 金晟, 等. 基于深度强化学习的室内视觉局部路径规划[J]. 智能系统学报, 2022, 17(5): 908−918.
ZHU Shaokai, MENG Qinghao, JIN Sheng, et al. Indoor visual local path planning based on deep reinforcement learning[J]. CAAI transactions on intelligent systems, 2022, 17(5): 908−918.

[6] 赵玉新, 杜登辉, 成小会, 等. 基于强化学习的海洋移动观测网络观测路径规划方法[J]. 智能系统学报, 2022, 17(1): 192−200.
ZHAO Yuxin, DU Denghui, CHENG Xiaohui, et al. Path planning for mobile ocean observation network based on reinforcement learning[J]. CAAI transactions on intelligent systems, 2022, 17(1): 192−200.

[7] 张晓明, 高士杰, 姚昌瑀, 等. 强化学习及其在机器人任务规划中的进展与分析[J]. 模式识别与人工智能, 2023, 36(10): 902−917.
ZHANG Xiaoming, GAO Shijie, YAO Changyu, et al. Reinforcement learning and its application in robot task planning: a survey[J]. Pattern recognition and artificial intelligence, 2023, 36(10): 902−917.

[8] 郭宪, 方勇纯. 仿生机器人运动步态控制: 强化学习方法综述[J]. 智能系统学报, 2020, 15(1): 152−159.
GUO Xian, FANG Yongchun. Locomotion gait control for bionic robots: a review of reinforcement learning methods[J]. CAAI transactions on intelligent systems, 2020, 15(1): 152−159.

[9] 乌兰, 刘全, 黄志刚, 等. 离线强化学习研究综述[J]. 计算机学报, 2025, 48(1): 156−187.
WU Lan, LIU Quan, HUANG Zhigang, et al. A review of research on offline reinforcement learning[J]. Chinese journal of computers, 2025, 48(1): 156−187.

[10] LEVINE S, KUMAR A, TUCKER G, et al. Offline reinforcement learning: tutorial, review, and perspectives on open problems[EB/OL]. (2020−05−04)[2024−10−12]. https://arxiv.org/pdf/2005.01643.

[11] 陈锶奇, 耿婕, 汪云飞, 等. 基于离线强化学习的研究综述[J]. 无线电通信技术, 2024, 50(5): 831−842.
CHEN Siqi, GENG Jie, WANG Yunfei, et al. Survey of research on offline reinforcement learning[J]. Radio communications technology, 2024, 50(5): 831−842.

[12] FUJIMOTO S, MEGER D, Precup D. Off-policy deep reinforcement learning without exploration[C]//International Conference on Machine Learning. Los Angeles: PMLR, 2019: 2052−2062.

[13] WU Yifan, TUCKER G, NACHUM O. Behavior regularized offline reinforcement learning[EB/OL]. (2019−11−26)[2024−10−12]. https://arxiv.org/abs/1911.11361v1.

[14] KUMAR A, FU J, SOH M, et al. Stabilizing off-policy Q-learning via bootstrapping error reduction[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019: 11784−11794.

[15] WANG Z, NOVIKOV A, ZOLNA K, et al. Critic regularized regression[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020: 7768−7778.

[16] FUJIMOTO S, VAN HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[C]//International Conference on Machine Learning. Stockholm: PMLR, 2018: 1587−1596.

[17] FU J, KUMAR A, NACHUM O, et al. D4RL: datasets for deep data-driven reinforcement learning[EB/OL]. (2021−02−06)[2024−10−12]. https://arxiv.org/abs/2004.07219v4.

[18] FIGUEIREDO PRUDENCIO R, MAXIMO M R O A, COLOMBINI E L. A survey on offline reinforcement learning: taxonomy, review, and open problems[J]. IEEE transactions on neural networks and learning systems, 2024, 35(8): 10237−10257.

[19] FUJIMOTO S, GU S S. A minimalist approach to offline reinforcement learning[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2021: 20132−20145.

[20] PENG Zhiyong, HAN Changlin, LIU Yadong, et al. Weighted policy constraints for offline reinforcement learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington DC: AAAI, 2023: 9435−9443.

[21] CHEN Xinyue, ZHOU Zijian, WANG Zheng, et al. Bail: best-action imitation learning for batch deep reinforcement learning[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020: 18353−18363.

[22] SIEGEL N Y, SPRINGENBERG J T, BERKENKAMP F, et al. Keep doing what worked: behavioral modelling priors for offline reinforcement learning[C]//International Conference on Learning Representations. [S. l.]: OpenReview.net, 2020: 1−21.

[23] ABDOLMALEKI A, SPRINGENBERG J T, TASSA Y, et al. Maximum a posteriori policy optimisation[C]//International Conference on Learning Representations. Vancouver: OpenReview.net, 2018: 1−23.

[24] BRANDFONBRENER D, WHITNEY W F, RANGAN-

ATH R, et al. Quantile filtered imitation learning[EB/OL]. (2021−12−02)[2024−10−12]. https://arxiv.org/abs/2112.00950v1.

[25] KOENKER R, HALLOCK K F. Quantile regression[J]. Journal of economic perspectives, 2001, 15(4): 143−156.

[26] AGARWAL R, SCHUURMANS D, NOROUZI M. An optimistic perspective on offline reinforcement learning[C]// International Conference on Machine Learning. [S. l.]: PMLR, 2020: 104–114.

[27] KOSTRIKOV I, NAIR A, LEVINE S. Offline reinforcement learning with implicit Q-learning[EB/OL]. (2021−10−12)[2024−10−12]. https://arxiv.org/abs/2110.06169v1.

[28] KUMAR A, ZHOU A, TUCKER G, et al. Conservative Q-learning for offline reinforcement learning[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020: 1179−1191.

[29] EMMONS S, EYSENBACH B, KOSTRIKOV I, et al. RvS: what is essential for offline RL via supervised learning? [EB/OL]. (2022−05−11)[2024−10−12]. https://arxiv.org/abs/2112.10751v2.

[30] CHEN Lili, LU K, RAJESWARAN A, et al. Decision transformer: reinforcement learning via sequence modeling[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2021: 15084−15097.

作者简介：

周娴玮，讲师，博士，主要研究方向为强化学习、机器人技术和多传感信息融合。E-mail: 20871147@qq.com。

王宇翔，硕士研究生，主要研究方向为深度强化学习和离线强化学习。E-mail: 2023024285@m.scnu.edu.cn。

余松森，教授，博士后，主要研究方向为智能感知与信息处理。主持国家自然科学基金面上项目 1 项、科技部星火计划面上项目 2 项、广东省基础与应用基础研究重点项目 1 项。参与制定广东省高端新型电子信息产业地方标准，获得发明专利授权 53 项，发表学术论文 40 余篇。E-mail: yss8109@163.com。