哈爾濱工業大學

算法设计与分析课程报告

班　号	2203101
学　号	2023140004
姓　名	阎发祥
专　业	计算机科学与技术
日　期	2024.5.1
授课教师	丁小欧

论文题目

Robust Task Representations for Offline Meta-Reinforcement Learning

via Contrastive Learning

作者：Haoqi Yuan, Zongqing Lu

刊物：International Conference on Machine Learning, 2022

出处：https://proceedings.mlr.press/v162/yuan22a.html

正文

1. 问题描述

离线元强化学习的范式中，离线数据的分布由行为策略和任务两部分共同决定，但是现有的算法无法辨别这些因素，使得任务表征对于行为策略的变化不够稳定。在 context-based 的设定下，OMRL（offline-metaRL）方法中，当行为策略与训练数据集中的任务高度相关时，context encoder 就很可能记住行为策略的特征。由此，在测试阶段，context encoder 就会因为行为策略的改变产生带有偏差的任务推断。之前的工作使用了对比学习来提升任务标准，但他们的学习目标是基于区分来自不同任务的轨迹，因此不能消除行为策略带来的影响。

2. 主要思想

本文提出了一种任务表征的对比学习框架，这种框架在面对训练与测试过程中的行为策略分布不匹配情况时有足够的鲁棒性。设计了一个双层 encoder 结构，使用共同信息最大化来规定任务表征学习，此外，引入了对比学习目标，通过生成模型和奖励随机化方法来近似负对样本的真实分布。

双层分别是：(1)从 one-step transition 中提取任务表征 (2)聚合这些表征

3.问题定义

3.1 马尔科夫决策过程

强化学习的任务被形式化为完全可观察的马尔可夫决策过程（MDP）。 MDP 被建模为元组 $M = (S, A, T, \rho, R, \gamma)$，其中 S 是状态空间， A 是动作空间， $T(s|s, a)$是转换环境的动态变化， $\rho(s)$是初始状态分布， $R(s, a)$是奖励函数， $\gamma \in [0, 1)$是对未来奖励进行贴现的因子。

3.2 离线元强化学习

任务服从一个 Mi 分布，不同的任务具有相同的行为空间与状态空间，但是奖励函数和 transition dynamics 是变化的，可以用 P(R,T)来表示任务分布。总共由 N 个训练任务。对于每一个任务i，一个离线数据表示为 $X_i = \{(s_{i,j}, a_{i,j}, r_{i,j}, s'_{i,j})\}_{j=1}^{K}$ 可以由任意的行为策略π_β^i收集。学习算法仅可以使用离线数据来训练元策略π_{meta}，不需要和任何环境交互。

在测试过程中，给定一个没见过的任务 $M \sim P(M)$，一个任意的探索（行为）策略收集一个 context c， $c = \{(s_j, a_j, r_j, s'_j)\}_{j=1}^{k}$ 之后智能体基于 c 执行任务适应，获取一个具体任务的策略 π_M，并在环境中进行评估。OMRL 的目标就是学习一个元策略在测试任务集上最大化期望回报。

context-based 元强化学习主要思想就是，不同任务具有一些相同的结构，并且任务的方差可以被一个表征所描述。因此 context-based OMRL 会用一个任务 encoder 来学习潜在空间参数z_t，来表示任务的信息。

4.算法描述和分析

4.1 双层任务编码

任务编码由 transition encoder E_{θ_1} 和聚合器 E_{θ_2} 构成，给定一个 context c，transition encoder 为所有 transition 提取隐藏信息

$$z_i = E_{\theta_1}(s_i, a_i, r_i, s'_i),$$

之后聚合器聚集所有隐层编码为一个任务表征

$$z = E_{\theta_2}(\{z_i\}_{i=1}^{k})$$

Q-函数$Q_\varphi(s,a,z)$和策略$\pi_\phi(a|s,z)$都以 z 为条件。

相比较于轨迹，transition 元组泄露更少的行为策略行为，更适合于 transition relabeling 以及 reward relabeling。transition encoder 结构就是简单的 MLP，聚合器利用 Q 函数以及使用离线 RL 算法的策略来训练

本能上，聚合器应该对具有更多任务信息的 transition 元组更为关注，所以使用了 self-attention 的方法

$$z = \sum_{j=1}^{k} softmax(\{MLP(z_i)\}_{i=1}^{k})_j \cdot z_j$$

4.2 对比任务表征学习

一个理想的鲁棒的任务表征应该独立于任务并且在各个行为策略下具有不变性。文章引入了互信息来衡量任务表征与任务之间的相互依赖性。相互信息衡量当观测到一个随机变量时,对另一个随机变量的不确定性减少的程度。（实际就是任务表示 z 和任务本身 M 之间的互信息 $I(z;M)$）作者希望最大化这个互信息,也就是最大程度地减少任务的不确定性,同时最小程度地保留与任务无关的信息。

$$maxI(z;M)=\mathbb{E}_{z,M}[log\frac{p(M|z)}{p(M)}]$$

互信息为 0 代表两个变量相互独立，越大两者之间依赖性越强，因此最大任务表示 z 和任务 M 的互信息就能学到最具代表性的特征。文章受到 InfoNCE（噪音对比估计），对比学习常用的一种损失函数的激励，推导出互信息的下界，并有以下理论

理论 1：M 是一个任务的集合，$|M|=N$ 且 $x=(s,a,r,s^{'}),z\sim p(z\mid x),h(x,z)=\frac{P(z\mid x)}{p(z)}$

推导出 $I(z,M)-log(N)\geq\mathbb{E}_{\mathcal{M},x,z}[log(\frac{h(x,z)}{\sum_{M*\in\mathcal{M}}h(x*,z)})]$

这个结构和对比学习常用的损失也是一样的。按照 InfoNCE，文章使用得分函数 $S(z*,z)$ 的指数来近似 h，这是两个样本潜在编码相似度之间的衡量。接着推导出一个容易处理下界的采样版本作为 transition encoder 的学习目标

$$\max_{\theta_1}\sum_{\substack{M_i\in\mathcal{M}\\x,x'\in X_i}}[log(\frac{exp(S(z,z'))}{\sum_{M*\in\mathcal{M}}exp(S(z,z*))})]$$

分段函数 S，实际上就是预先相似度。

$$similarity=cos(\theta)=\frac{A\cdot B}{||A||||B||}$$

(x,x') 就是正例，$(x,x*)$ 是负例。为了最大化正例的分数，transition encoder 一个抽取相同任务的 transition 的共享特征，并学习到最关键的 rewards 和 state transition 的差异。

4.3 负例的生成

这一部分是论文的关键。在完全离线情况下，每个任务的数据量有限，不足以准确学习该任务的奖励函数 R 或者状态转移函数。即使能够学习到 R 和 T，也需要当前状态 s 和动作 a 在所有任务中都有覆盖，才能依赖 R 和 T 生成其他任务的(r, s')，通常很难满足。

在离线的设定下，我们需要适应更高维度的转换模型，但是每个任务的数据集往往比

较小，使得奖励模型和转换模型容易过拟合。当不同任务之间 状态-行为对重叠较小的情况下，利用利用单独学习的模型进行重新标注可能不准确，使得对比学习更加精确。

a. Fidelity

为了确保 task encoder 在真实的转化元组上进行学习，生成的负例应该近似真实分布。给定状态 s 和动作 a，负样本 $(r, s^{'})$ 的生成分布应该正比于所有任务 M 的状态转移函数在满足 $R(s,a)=r$ 的约束下的加权和，就是根据所有人物的奖励函数 R 和转移函数 T，生成符合这些任务分布的负样本。

$$p(r, s'|s, a) \propto \mathbb{E}_{M \backsim P(M)}[T(s'|s, a)\mathbf{1}\{R(s, a) = r\}]$$

b. diversity

对比学习的性能表现随着负例多样性增加而提高

4.3.1 生成式建模

将奖励和 transition 模型与跨任务的更大数据集进行拟合可以减少预测误差，因此在训练数据集的并集上训练生成式模型来近似离线数据的分布

$$P_{\{x_i\}}(r, s^{'}|s, a)$$

使用 conditional VAE（CVAE），CVAE 由一个生成器 $p_\eta\left(r, s'|s, a, z\right)$ 和一个概率 encoder $q_\omega\left(z|s, a, r, s'\right)$ 构成。隐向量 z 描述了预测的不确定性因素，遵循先验高斯分布 p(z)。

CVAE 最小化损失

$$\begin{aligned}\mathcal{L}_{CVAE} = &-\mathbb{E}_{(s,a,r,s')\in\{X_i\}}[\mathbb{E}_{q_\omega}[log\, p_\xi(r, s'|s, a, z)] \\ &- KL[q_\omega(z|s, a, r, s')||p(z)]]\end{aligned}$$

第一项：重构损失，使 decoder 可以根据 z 恢复样本

第二项：KL 散度，使 latent code z 的分布尽可能接近先验高斯分布，作为正则项

在训练过程中：

1) Encoder 接收 (s, a, r, s') 作为输入,输出 z 的分布 $q_\omega(z)$。

2) 从 $q_\omega(z)$ 中采样 z。

3) Decoder 接收 (s, a, z) 作为输入,预测 (r, s') 的分布 $p_\xi(r, s')$。

4) 计算重构损失和 KL 散度,更新网络的参数 ω 和 ξ。

在训练结束后,我们可以只使用 Decoder pξ(r,s'|s,a,z)来生成负样本,输入不同的(s,a)和采样不同的 z 即可获得样本分布。

4.3.2 奖励随机化

在各任务之间状态-行为对重叠较小的情况下，CVAE 可能坍塌成成确定性预测模型导致负例多样性降低。当任务仅有奖励函数不同的情况下，通过给奖励添加随机噪音来生成负例

$r^* = r + v$， v 服从噪声分布 $p(v)$

算法总结如下图

Algorithm 1. Meta Training

Input: Datasets $\{X_i\}_{i=1}^N$; OMRL models $E_{\theta_1}, E_{\theta_2}, Q_\psi, \pi_\phi$
A. If use generative modeling, pre-train CVAE:
Initialize CVAE q_ω, p_ξ
repeat
  Update ω, ξ to minimize Eq. (10).
until Done
B. Train the transition encoder:
repeat
  Sample a task M and two transition tuples x, x'
  $z = E_{\theta_1}(x), z' = E_{\theta_1}(x')$
  for $M^* \in \mathcal{M}$ **do**
    if use generative modeling **then**
      Sample x^* from CVAE
    else if use reward randomization **then**
      Add noise to the reward to get x^*
    end if
    $z^* = E_{\theta_1}(x^*)$
  end for
  Compute Eq. (8)
  Update θ_1 to maximize Eq. (8)
until Done
C. Train the policy:
repeat
  Sample a task dataset X and a context c
  $z = E_{\theta_2}(E_{\theta_1}(c))$
  Augment the states in X with z
  Update θ_2, ψ, ϕ with offline RL algorithms on X
until Done

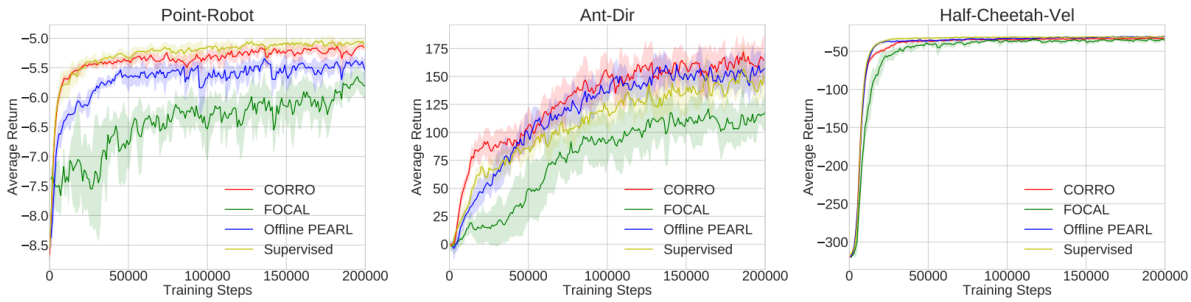
测试阶段

Algorithm 2. Meta Test

Input: Trained models $E_{\theta_1}, E_{\theta_2}, Q_\psi, \pi_\phi$
Sample a task M
Collect a context trajectory c with an arbitrary policy
$z = E_{\theta_2}(E_{\theta_1}(c))$
repeat
  Observe s, execute $a \sim \pi_\phi(a|s, z)$, get r
until Environment terminates

4.4 算法实例

一个具体的算法例子是 CORRO 框架中的任务编码器，使用双层结构：第一层为转换编码器（transition encoder），用于从单步转换元组中提取潜在表示；第二层为聚合器（aggregator），用于将所有潜在代码聚合成任务表示。这种结构可以有效地从转换中学习任务信息，同时通过对比学习最大化任务表示和任务之间的互信息，以增强任务表示的鲁棒性。实验结果如下图所示，文章的实验结果表明，提出的方法（CORRO）在多种离线元强化学习基准测试中，尤其是在行为策略出现分布外情况时，表现优于现有的基于上下文的 OMRL 方法。该方法可以在完全离线的数据集上学习鲁棒的任务表示，能够从任务的转换分布中区分任务，而不受行为策略的影响。



5 讨论和分析

5.1 相关领域问题分析

文章涉及的相关领域包括：离线强化学习（Offline Reinforcement Learning）、元强化学习（Meta-Reinforcement Learning）、对比学习（Contrastive Learning）。离线强化学习关注如何从预先收集的数据中学习策略，元强化学习关注如何快速适应新任务，对比学习用于自监督表示学习中，通过区分正负样本对来提取有用的特征。

- 离线强化学习的核心问题是如何在没有进一步与环境互动的情况下，仅利用预先收集的数据来学习策略。这涉及到数据效率问题和策略的泛化能力。文中通过建立批量强化学习（Batch RL）的方法，使用 Q-learning 或重要性采样来重新利用已有数据。还提到了利用对比学习来增强任务表示的稳健性，以解决行为策略在训练和测试时分布不匹配的问题。
- 在元强化学习中，任务是快速适应新任务，关键是如何在多任务分布上训练以达到快速泛化。文中通过上下文编码器学习任务表示。采用优化基方法和上下文基方法，上下文基方法将任务视为状态的不可观测部分，并从历史轨迹中编码任务信息。提出的 CORRO 框架利用对比学习提高任务表示的鲁棒性，尤其是对行为策略变化的鲁棒性。
- 在对比学习中，负样本的选择对于学习有效的表示至关重要。然而，传统方法中常常难以准确选择合适的负样本。在实际场景中，正样本与负样本的比例可能极不平衡，这会导致模型倾向于学习更常见的类别，而忽视罕见的类别。同时学习到的表示可能对于训

练数据有效，但在面对新的、未见过的数据时，泛化能力不足。采用主动挖掘负样本的方法，如使用对抗性生成网络（GAN）来生成更具挑战性的负样本，或者使用任务相关的标准来选择负样本可以优化问题。使用过采样、欠采样或者集成学习等技术来调整正负样本的比例，可以缓解数据不平衡带来的问题。结合多任务学习的思想，同时学习任务特定的表示和任务无关的通用表示，能够提高表示的泛化能力。

这三个领域都关注于如何从有限的数据中学习有效的策略，同时解决数据效率和泛化能力问题。CORRO 框架通过结合对比学习和元学习的方法，尤其强调在离线设置中提高任务表示的稳健性和泛化能力，这在传统方法中是一个较大的挑战。通过实验验证，CORRO 在多种离线元强化学习基准上表现优于现有方法，特别是在行为策略为分布外时的表现。

5.2 算法问题与改进

文中提到的现有算法面临的主要问题是如何从离线数据中提取稳健的任务表示，尤其是在行为策略和任务目标之间存在分布偏差时。这些问题主要表现在：

- 任务表示的稳健性不足：现有的方法依赖于行为策略生成的数据，容易受到行为策略变化的影响，从而导致在新任务上泛化能力弱。
- 对负对样本选择的方法局限性：传统对比学习中负对样本的选择往往随机，没有充分考虑到任务相关性，可能影响任务表示的学习效率和质量。
- 对不同任务间差异性编码不足：现有的算法可能不足以区分不同任务间的微妙差异，特别是在任务相似但不完全相同的情况下。

为了解决上述问题，我们可以考虑在原有算法的基础上进行改进。通过引入任务相关的负对样本挖掘，可以有效增强任务表示的区分度和鲁棒性。这种方法直接对抗了现有方法在任务表示上的主要弱点，即在面对行为策略变化时的脆弱性。同时，使用图神经网络整合多任务信息，能够在不同任务之间传递有用的学习信息，提高表示的泛化能力。此外，动态任务编码调整机制使算法具有更高的灵活性和适应性，能够根据实际表现调整学习策略，进一步优化学习效果。这些特性使得该框架不仅能够在标准的离线元强化学习设置中表现良好，也能在行为策略高度变化的复杂环境中维持高效的学习性能。