arXiv:2304.14126v1 [cs.AI] 27 Apr 2023

从示范中推断多目标多智能体决策中的偏好

预印本

Junlin Lu *计算机科
学学院 戈尔韦大学
J.Lu5@nuigalway.ie

2023年4月28日

摘要

在多目标决策问题中，量化不同目标的数值偏好颇具挑战。然而，用户的示范通常是可以获取的。我们提出了一种算法，用于从最优或近最优示范中推断线性偏好权重。该算法在三个环境中与两种基线方法进行了评估。实证结果表明，与基线算法相比，该算法在时间需求和推断偏好的准确性方面均有显著提升。在未来的工作中，我们计划在多智能体系统中评估该算法的有效性，其中一个智能体将利用我们的偏好推断算法来推断对手的偏好。

关键词 多目标 强化学习 · 偏好推断 · 动态权重多目标智能体 .多智能体系统 · 对手建模

1 引言

在多目标决策过程中，智能体会针对不同目标接收奖励向量。效用函数通过将奖励向量评估为标量效用，用于在相互竞争的目标之间进行权衡。在现有文献中，获取标量效用最常用的方法是线性标量化 [Castelletti 等，2013；Khamis 和 Gomaa，2014；Ferreira 等，2017；Lu 等， 2022]。在线性标量化中，奖励的权重被称为偏好。然而，对用户而言，给出精确的数值偏好并不总是直观的。例如，考虑这样一种情况：投资组合经理根据最小化风险和最大化利润的权重来选择股票。他/她可能希望赋予最大化利润更高的权重，但很难确定一个具体数值。偏好中的微小误差可能导致截然不同的策略，进而产生次优解。然而，尽管难以用数值明确表示偏好，用户通常可以示范其偏好。因此，在解决此类问题时，使用示范的偏好推断（PI）方法会很有帮助。

此外，在多智能体多目标决策过程中，如果某个智能体能够使用偏好推断机制来推断其他智能体的偏好，就能获得有关其他智能体的信息。了解他人的偏好可以让这些“明智”的智能体占据优势。

2 偏好推断

PI 是推断一组权重，这些权重用于在多目标决策过程中将奖励向量标量化。这与逆强化学习（IRL）类似，后者通过寻找一组能够将与状态相关的特征标量化的线性参数来推断奖励函数。这里有两个假设：

*Junlin Lu 的研究得到爱尔兰政府研究生奖学金（GOIPG/2022/2140）的资助。

- 观察到的轨迹来自最优策略或近最优策略。这是现有文献中对 IRL 和 PI 都广泛接受的假设 [Ng 等，2000；Ziebart 等，2008；Wulfmeier 等， 2015] 以及 PI [Takayama 和 Arai， 2022]。
- 基于第一个假设，在给定最优或近最优策略的情况下，平均奖励轨迹完全由偏好和环境转移决定。

当我们获得一个被假定处于帕累托最优集（POS）上的点，且该点基于效用函数的某些未知权重生成，并希望确切了解这些权重时，便产生了偏好推断问题。对于更现实的场景，我们还考虑了接近帕累托最优集上各点的被支配点，这些点被称为次优策略，用于测试偏好推断算法。通过向数据中加入次优噪声，我们确保推断模型能够适应次优奖励轨迹。

我们首先使用动态权重强化学习（DWRL）智能体生成行为轨迹 [Källström 和 Heintz， 2019]。随后，我们提出基于动态权重的偏好推断（DWPI）算法，以推断智能体针对不同行为轨迹的偏好。DWPI 算法的训练过程如图 1 所示。DWRL 智能体将偏好向量作为状态的一部分，从而能够改变其行为模式

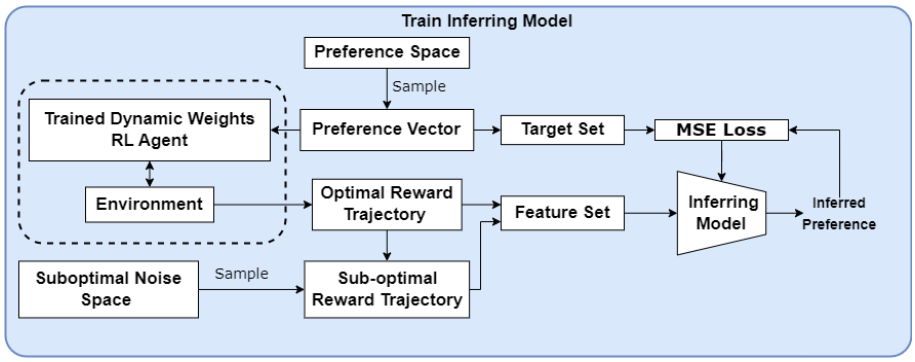

图 1：动态权重偏好推断算法训练阶段

在运行时通过改变偏好来实现。给定经过训练的动态权重强化学习智能体与环境之间的交互，可生成一组最优奖励轨迹。通过添加随机采样的次优噪声，对轨迹集进行扩充，使其成为推断模型的训练集。在监督学习范式下，通过输入奖励轨迹并预测相应偏好，对推断模型进行训练。

算法 1动态权重偏好推断算法

初始化推断模型 $\mathcal{I}$、次优噪声空间 $\mathcal{SN}$、环境 $\mathcal{E}$ 和偏好空间 Ω
加载经过训练的动态权重强化学习智能体 AG
初始化特征集 X、目标集 Y
当 条目数量不足时 X **执行**
  从 Ω 中采样偏好向量 $\boldsymbol{\omega}$
  AG 与 $\boldsymbol{\omega}$进行一回合交互，生成奖励轨迹 $\boldsymbol{\tau_r}$
  从 $\mathcal{SN}$ 中采样噪声向量 $\boldsymbol{\delta}$
  将带噪奖励轨迹 $\boldsymbol{\tau_r} + \boldsymbol{\delta}$ 存储在 X 中，将 $\boldsymbol{\omega}$ 存储在 Y 中
结束 while 循环
当 $\mathcal{I}$ 不收敛**时**
  从 X 中抽取批次以训练推断模型，损失 $\mathcal{L} = ||\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}||$
当结束

我们在三个环境中评估我们的算法：凸深海寻宝[Mannion 等人， 2017], 交通[Källström 和 Heintz， 2019], 以及物品收集[Källström 和 Heintz， 2019] 并将其与两种基准方法进行比较：投影方法（PM） [Ikenaga 和 Arai， 2018] 以及乘法权重学徒学习（MWAL） [Takayama 和 Arai， 2022]。我们的方法在时间效率（图 2）和偏好推断（PI）性能（表 1）方面均优于基线方法。实验使用 Python 3.9 和 TensorFlow 2.3.0 实现，并在配备第 11 代 Intel(R) Core(TM) i7-1165G7 2.80GHz CPU 的机器上运行。

结果　　评估结果表明，DWPI 算法在推断准确率方面表现良好　　.

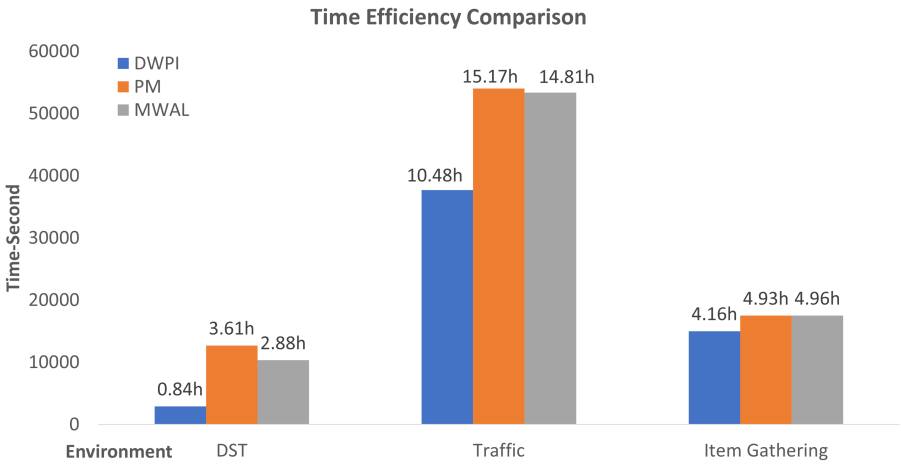

图 2：时间效率比较

表 1：性能提升

环境		交通		物品收集	
KL散度		PM	MWAL	PM	MWAL
最优演示	DWPI	90.8%↑	99.56%↑	98.01%↑	99.13%↑
次优演示	DWPI	89.55%↑	99.53%↑	96.89%↑	99.25%↑
均方误差		PM	MWAL	PM	MWAL
最优演示	DWPI	97.7%↑	99.8%↑	85.91%↑	98.1%↑
次优演示	DWPI	94.13%↑	99.83%↑	83.81%↑	98.9%↑
实用性		PM	MWAL	PM	MWAL
最优演示	DWPI	60.56%↑	98.93%↑	90.67%↑	82.62%↑
次优演示	DWPI	71.87%↑	90.60%↑	99.85%↑	94.50%↑

3 对手偏好建模

在未来的工作中，我们将利用 DWPI 算法，使智能体能够在多智能体环境中了解其他智能体的偏好，从而获得相对于它们的优势。
将在两个环境中进行评估。第一个是 Wolfpack环境 [Leibo 等人， 2017], 其中两个捕食者智能体试图捕获随机移动的猎物。如果在两个捕食者之间的曼哈顿距离为 $dist \leq 3$ 时捕获猎物，则根据 $dist > 3$ 将其判定为合作或竞争。

第二个环境是基于 [Källström 和 Heintz， 2019] 修改的多智能体物品收集环境。两个强化学习智能体在网格世界中移动，收集随机分布的三种不同颜色的方块，即绿色、红色和黄色。每个智能体对不同颜色的方块都有一个偏好权重向量。

将在每个环境中进行两组实验。第一组实验在两个普通智能体之间进行，而第二组实验在一个普通智能体和一个能够推断另一方偏好的更智能的智能体之间进行。将分析智能体的表现，以检验推断机制是否能帮助更智能的智能体在博弈过程中获得优势。

4 结论

我们提出 DWPI 算法，以从多目标决策过程中学习者的示范中推断偏好。我们进一步评估该算法在提升多智能体系统中智能体表现方面的效用。借助我们的 PI 模型，智能体能够更好地了解其对手，从而在目标任务上取得更好的表现。未来，我们希望评估多智能体系统中的对手偏好建模效果。将 DWPI 算法扩展到推断非线性偏好，也是我们感兴趣的研究方向。

参考文献

Andrea Castelletti、Francesca Pianosi 和 Marcello Restelli。用于水资源系统运行的多目标强化学习方法：单次运行中的帕累托前沿近似。*WaterResources Research*，49(6)：3476–3486，2013年。Mohamed A Khamis 和 Walid Gomaa。基于协作多智能体框架、采用混合探索的自适应多目标强化学习，用于交通信号控制。*Engineering Applications of Artificial Intelligence*，29：134–151，2014年。Paulo Victor R Ferreira、Randy Paffenroth、Alexander M Wyglinski、Timothy M Hackett、Sven G Bilén、Richard C Reinhart 和 Dale J Mortensen。用于认知空间通信的基于多目标强化学习的深度神经网络。载于*2017*年认知通信航空航天应用研讨会（*CCAA*），第1–8页。IEEE，2017年。Junlin Lu、Patrick Mannion 和 Karl Mason。用于住宅电器调度的多目标多智能体深度强化学习方法。*IET Smart Grid*，2022年。Andrew Y Ng、Stuart J Russell 等。逆强化学习算法。载于*Icml*，第1卷，第2页，2000年。Brian D Ziebart、Andrew L Maas、J Andrew Bagnell、Anind K Dey 等。最大熵逆强化学习。载于*Aaai*，第8卷，第1433–1438页。美国伊利诺伊州芝加哥，2008年。Markus Wulfmeier、Peter Ondruska 和 Ingmar Posner。深度逆强化学习。*CoRR，abs/1507.04888*，2015年。Naoya Takayama 和 Sachiyo Arai。用于目标权重估计的多目标深度逆强化学习。*Artificial Life and Robotics*，第1–9页，2022年。Johan Källström 和 Fredrik Heintz。利用多目标强化学习在基于智能体的仿真中实现可调动态。载于*2019*年*5*月*13*日至*14*日在加拿大蒙特利尔举行的 *AAMAS* 自适应与学习智能体研讨会（*ALA-19*），第1–7页，2019年。Patrick Mannion、Sam Devlin、Karl Mason、Jim Duggan 和 Enda Howley。多目标强化学习中奖励变换下的策略不变性。*Neurocomputing*，263：60–73，2017年。Akiko Ikenaga 和 Sachiyo Arai。用于多目标序贯优化中偏好 elicitation 的逆强化学习方法。载于*2018*年 *IEEE* 国际智能体会议（*ICA*），第 117–118 页。IEEE，2018 年。Joel Z Leibo 、 Vinicius Zambaldi 、 Marc Lanctot、Janusz Marecki 和 Thore Graepel。序贯社会困境中的多智能体强化学习。*arXiv* 预印本 *arXiv:1702.03037*，2017年。