强化学习中的异策略评估

作者：writer 来源：中国科学院

本文原地址：https://www.iikx.com/news/progress/5348.html

强化学习中的异策略评估。近日，Frontiers of Computer Science (FCS)期刊发表来自美国谷歌大脑的研究科学家Lihong Li的观点文章——A perspective on off-policy evaluation in reinforcement learning，探讨强化学习的异策略评估，该评估给出了一种廉价而安全的评价强化学习算法的途径，有望能够释放强化学习的力量。

1.背景

强化学习(RL)的目标是构建一个自主智能体，通过与未知的外部环境交互，该智能体学习使效用函数最大化的行为序列。它是一种非常通用的学习范式，可以用于对各种各样问题的建模，比如游戏、机器人、自动驾驶、人机交互、推荐、健康护理等等。近些年，得益于深度学习和计算能力的进步，强化学习取得了很大的成功，AlphaGo/AlphaZero就是一个著名的例子。这些令人惊叹的成果，激发了人们应用强化学习以解决现实问题的兴趣。

在强化学习中，智能体策略的好坏，往往通过平均回报来度量。如果智能体所在的环境是可模拟的，例如计算机游戏环境，那么可以通过实际运行这个策略，来获得评估结果。但是，对于多数现实场景，例如自动驾驶和医药治疗应用，直接在实际环境中运行新策略的成本昂贵、风险巨大，甚至涉及道德问题。因此，通常的实践中，常常会构造环境模拟器，用于策略的评估。但是，构建一个高精度的模拟器的工作，往往比寻找最优策略本身还要困难(例如考虑如何构建一个能够覆盖所有医疗状况的模拟病人)。因此，强化学习实践者经常痛苦地发现他们处于一个死局中：为了能使用一个策略，必须先通过评估验证策略的质量合格，但对于策略而言，其唯一可靠的评估方法却是去使用这一策略!

2.问题

以上阐述的挑战引发了对异策略评估的需求，即对一个策略(目标策略)的评估只使用由另一个策略(行为策略)执行产生的历史数据，而并不实际运行目标策略。这个问题可能听起来很简单，但事实上却是强化学习过去数十年中最为关键和基础的研究主题之一。

我们可以通过与监督学习(SL)做对比来了解存在的挑战。 以构建垃圾邮件检测器为例，垃圾邮件检测结果的评价很直接：给出一个垃圾邮件分类器，可以用标记数据来测量它的准确率(或是其他指标)，准确率越高分类器就越好。而强化学习面临的情况则复杂得多。强化学习的数据通常是轨迹的形式，组织成状态—动作—回报元组的序列，一个时刻的状态由序列中前一时刻的动作决定。因此，如果策略在某个时刻偏离了轨迹数据(即选择了一个与数据记录所不同的动作)，那么所有未来的状态和回报都可能改变，但新的状态并没有出现在数据中。换句话说，与监督学习

不同，强化学习的数据仅能为策略评价提供部分信息。因此，异策略评估需要利用反事实推理，以回答如果-会怎样的问题，这与因果推断密切相关。

(1)上下文赌博机情况(译者注：即单步决策情况)

异策略评估在强化学习任务的一个重要的子类，即上下文赌博机(contextual bandits)中较容易实现。在这样的环境中，智能体的行为不会影响未来的状态，但数据中仅包含行为的回报数据，因此仍然需要进行反事实推理。上下文赌博机可以用于很多重要应用的建模，例如推荐、广告和网页搜索等，在这些应用中回报可能取决于用户的点击、视频浏览的时间或者取得的收入。

一类基于逆倾向评分(inverse propensity scoring ,IPS)的强大方法在实践中被证明有效。它们使用重要性取样修正观测数据(行为策略采样数据)与期望但未观察数据(目标策略所需数据)之间分布的不一致。目标策略的评价通过对回报数据的重要性加权平均来计算。在宽松的假设条件下，IPS的估计是无偏的，并且随着数据的增加趋向目标策略的真实值。IPS方法的主要缺陷在于其估计的方差较大。随后产生了许多降低其方差的方法，或许以增加少许的偏差为代价，以获得一个更加准确的估计。

(2)一般强化学习情况(译者注：即多步决策情况)

IPS 方法可以延伸到更一般的情况中，即智能体的行为会影响未来的状态。理论上，仅需要将重要性采样应用至整个轨迹即可。但遗憾的是，这样的方法会使估计的方差随着轨迹长度的增长指数爆炸，这一现象称为视域灾难(the curse of horizon)。因此，在实际中这类方法未被广泛使用。

最近，又有一类新的方法出现，仅计算状态上的重要性权重，而不是轨迹的权重，因此避免了对于轨迹长度的直接的依赖。其首个算法就展示出了良好的前景，而更强的算法也正在发展中。

3.结论

异策略评估使上下文赌博机模型成功在网页应用中使用，并且在推动赌博机模型实用化上起到关键作用。在一般的强化场景也可以采用同样的思路。可靠的异策略评估有望能够释放强化学习的力量。它给出了一种廉价而安全的评价强化学习算法的途径。

还有很多问题值得进一步研究，在此列举一二。首先，我们对所面临问题的统计本质还缺乏理论理解，尤其是对于一般强化学习的情况。其次，大多数本领域发展的通用算法可视为在偏差-方差上寻找平衡。而与探寻通用技术不同，在具体应用中通过发现有效的结构，例如减少有效动作数量，可以取得降低方差的效果。第三，我们的讨论仅集中在异策略评估，而更具挑战的是其下一步——异策略优化，即在行为策略收集的历史数据上优化策略。

作者简介：

Lihong LI：美国谷歌大脑的研究科学家。他的主要研究领域是强化学习，包括上下文赌博机和其他人工智能相关方向。他的工作已经应用于推荐信、广告、网络搜索和对话系统，并在ICML、AISTATS和WSDM获得最佳论文奖。他在主要的AI/ML会议(如AAAI、ICLR、ICML、IJCAI和NIPS/NEURIPS)中担任领域主席或高级项目委员会成员。

Frontiers of Computer Science(FCS)是由高等教育出版社与北京航空航天大学共同主办，Springer海外发行的英文国际期刊，双月出版。主编为南京大学周志华教授;共同主编为北京航空航天大学熊璋教授。主要刊登计算机科学领域具有创新性的研究成果。涉及领域包括(但不限于)体系结构，软件，人工智能，理论计算机科学，网络及通信，信息系统，多媒体及图像，信息安全，以及交叉领域等。文章类型包括：研究论文、综述及短文，并设有特色专栏：Perspective、优秀青年科学家论坛。本刊已被SCI、Ei、DBLP、INSPEC、SCOPUS和中国科学引文数据库(CSCD)核心库等收录，为CCF推荐期刊;两次入选中国科技期刊国际影响力提升计划;入选第4届中国国际化精品科技期刊。最新影响因子1.105。