从延迟奖励中学习

FrankZhou-jun*

2019 年 10 月 22 日

摘要

在行为生态学中，随机动态规划作为一种计算动物最优行为决策的方法，得到普遍使用，但是动物是怎样从他们的经历之中学习到最优策略的呢？这篇文章针对学习高效行为的可能方法做了一个系统的分析。

首先讨论的是，它遵守动物应该学习最优策略的最优假设，，即便动物可能不遵守，接下来讨论的是马尔科夫决策过程，这是一个普遍的模型，及动物与环境之间的交互，通过动态规划决定最优策略是最普遍的方法，将在后文提及。动物进行这类计算是不合理的。

可是，这里有许多方法，可以组成动态规划计算的方法，可能是动物学习的合理模型，特别指出，蒙特卡洛 (Monte-Carlo) 方法可以直接学习动作的最优值 (或规范成本)。不需要给动物模拟环境。也不需要长时间记住状态和动作。这个学习方法给出了证明，这类学习方法也可能用于层级策略。重观以前的学习方法，一些简单的方法没有给出证据就给了出来。本文描述了部分学习方法的验证过程。

*研究方向：信号处理，机械故障诊断，深度学习，强化学习，邮箱:zhoujun14@yeah.net

目录

1 绪论 **2**

1.1 经典型条件反射和操作性条件反射 (也叫工具性条件反射) 2

1.2 最优性论证 3

2 致谢 **3**

1 绪论

以奖励的方式学习动作是一个智能的象征，例如一条狗，当它做出与命令一致的动作就给予奖励，以奖励的方式训练。我们普遍赞同动物可以学会得到奖励值，而避免惩罚，在心理学试验中，动物的智力已经得到广泛的研究。但是在学习方式的研究被忽略了。

有关如何从奖励和惩罚信号中学习，本文将会提出一个总的计算方法，可能会用于现存的研究动物学习的研究工作中，也可以用于其他类型的学习问题。这篇文章的目的并不是使用特定的计算模型去解释特定心理学实验的结果，而是规范地提出一些可以被动物使用来优化动作的算法，这些算法可能在自适应控制系统和人工智能方面具有应用前景。

这章节我们将讨论怎样研究动物学习和学习计算理论的要求是什么。

1.1 经典型条件反射和操作性条件反射 (也叫工具性条件反射)

现有许多有关动物学习试验研究的文献，我不会做综合性的评价，相反，我将描述试验研究的重要目的，现象研究的特点，以及主要结论。

人们对状态和联想学习研究具有很长的一段历史，可以用过马尔科夫进行描述，为了研究动物的学习能力，我们做了一个人工环境，实验者可以控制事件或不测事件的发生，最初的人工环境是斯金纳箱，动物在里面可能会面临刺激，例如蜂鸣声，或者是白炽灯光。动物可能会做出一些动作，例如在老鼠笼中按压摇杆，在鸽子笼中啄光，给动物提供一个强化刺激。在行为学领域，一个正向的刺激可以提高之前的反应的能力，例如给饥饿的动物一点点食物或者给饥渴的动物一点水。相反的，一个负向的刺激，例如，电击，可以减少之前反应的能力。动物环境可以被自动执行很长一段时间，根据呈现的刺激和动物的反应投递出强化刺激。事件和不确定事件的规定就是根据著名强化表来的。使用的两种基本的试验程序被使用，工具性条件反射和经典型条件反射表

在操作性表中，动物得到强化是取决于它做了什么，操作性学习是学习执行可以得到奖励的动作而避免得到惩罚：动物以一定的方式学习行为，因为在那种方式下可以带来正向强化，在动物中，操作性条件反射的自适应作用非常清楚，如果蓝雀可以学会仔细观察放置的鸟桌，在冬天他可以获得更多的食物。在经典条件反射（巴浦洛夫反射）试验，动物受到一系列事件然后强化刺激。在这些事件中强化刺激是非常偶然的，在动物的行为中却不是偶然的，例如，老鼠可能遇到灯光，然后给它电击，而不考虑它做出什么行为。这类实验是更好的，因为这些事件和和强化刺激是可以通过实验者控制的，但是动物行动可能不受控制。

古典型条件反射实验可能需要动物不通过之前的学习，本能的对某种刺激进行反应：一个人遇到针尖时会本能收手，一条狗看见大量的食物会流口水，引起反应的这些刺激称为无条件刺激（unconditional stimulus，US），如果动物放在其他刺激的环境，则称为条件刺激（condition stimulus，CS），一般发生在无条件刺激之前，因此条件刺激的发生一般伴随着无条件刺激发生，动物只能在条件刺激之后做出反应。让动物做出条件刺激的结果就像其网络中无条件刺激的出现的结果一样。古典条件反射和操作性条件反射是否为两种类型的受到很大争议，在试图通过试验来说明这个问题时变得更加复杂和困难。在此，我关心的是学

习算法而不是动物实验，因此我将会对动物实验的试验证据进行一个简短的解释，为我后面研究的学习算法提供一个支撑。

迈金托时详细的讨论了古典型条件反射和操作性条件反射。首先，他试图将经典型条件反射作为操作性条件反射解释：在期望食物时，可能狗不是学习分泌唾液，因为它发现，即使分泌唾液食物难道就会更加可口吗？迈进托时认为通过这种方式，古典型条件反射不能被认为是操作性条件反射。

迈克托时惊喜的发现到许多操作性条件反射明显地可以解释为古典型天界反射。在操作性条件反射实验中，动物根据条件刺激学习一些行为的过程中，动物不可避免的要观测条件刺激和自己行为产生何种奖励之间的联系。由动物自己引起的这种关系可能会导致古典型条件反应：如果古典型条件反应和操作性反应一样，那么每一次的动作将会加强条件刺激和奖励之间的联系，因此加强了条件刺激的持续。正反馈学习过程：动物的反应越稳定，它所观测到的关系就越好，同时它所观测到的关系就越好，动物的反应就越稳定。

但是，迈克托时称并不是所有的操作性学习可以用这种方式解释，一个直接明显的理由是动物可以根据相同的刺激学习不同的行为来获得相同的奖励。

迈克托时表示没在操作性条件反射实验中会受到古典型条件反射的影响，反之亦然。为了进行操作性学习，动作必须要做出一个反应，或者一系列反应来获得奖励，但是为什么动物可以产生第一个反应？操作性学习的组成：通过尝试 $\rightarrow$ 重复之前的成功经历。而第一次成功的实现有不同的解释。一种可能是动物随机地对环境进行探测，但是这并不是一个完整的解释，一个合理假设是当某种事件类型与环境相关时，动物本能产生的合适动作 (古典型条件反射)。一个无伤大雅的解释是这个合适的动作是古典型条件反射产生的，然后通过操作性学习进行微调。在古典型反射和操作性条件反射之间的问题是：动物具有什么先天性的本能，这些本能是以何种方式帮助学习？

条件反射理论试图详细地解释动物的行为，例如，动物做出的反应和强化刺激物之间的时间间隔是如何影响反应学习的速率，但在更多的自然条件下，这些研究都不能很好用于解释或者预测动物的行为，当操作性条件反射构建的模型，刺激，反应，和强化之间的关系变得越来越复杂时而不能预测。很明明显操作性条件反射可以使动物学习已到达得到自己所需的目标，但是在实际更多的自然条件中，条件反射理论不能量化的预测学习的结果。则问题是大多条件反射理论倾向解释某种实验类型下得到的结果，而不是预测在某种行为上的学习效果。

1.2 最优性论证

行为生态学家开始从不同的方向解释动物的行为，他们认为动物要存活和繁衍下去，则需要有效的行为方式，应该有来自选择的压力，使动物采取某种行为策略确保最大化的成功繁衍。

2 致谢

感谢在实验室度过的两年时光，老师无论在学术还是人生的指导上都对我起到了很大的帮助；师兄师姐小伙伴们的鼓励支持和陪伴是我坚持下去的动力。