# 强化学习经典入门书的读书笔记系列--第一篇

2017-04-15 小吕 神经网络与强化学习

这是最近读sutton的“reinforcement learning - An introduction”的读书笔记，其实大多数更像是按照自己理解写的中文翻译，可能语言有时没有那么严谨，主观因素多一点，更像是一种和自己的对话。希望各位看官多多包涵，如果想真正理解一些，最好对照着英文原本看，也许能看出自己的想法。

这次第一篇就写第一章。第一章是概述，更多的是从宏观上讲强化学习的一些概念和思想，虽然概括性较强，但也还是有很多值得细读的点，在下文一一道来。

１、强化学习中的基本元素：

policy --相当与环境和动作之间的一个映射，某种环境下最应该做什么动作呢？这个是由policy决定的。policy的所有可能组成一个policy空间，强化学习的目的，就是在这个巨大的空间中，学习到某一种最优的policy。

reward function -- 也可以看成是一个映射，关于当前的动作，或者当前环境和当前动作的pair的好不好的一个评价。属于立即评价，只考虑当前这一步的好坏。

value function -- 和上面的reward function对比着看，这一步考虑的是当前环境状态的长远优势，也就是以当前状态为起点，以后的多个时间点之后的各个状态的reward之和。如何更好的估计这个值，是几乎所有增强学习问题的解决重点和难点。这个也是如何评定一个policy好坏的标准。也是把增加学习和evoluation method （例如遗传算法）区别开的地方。

model of environment -- 对环境的建模。这个模型有点预测环境的走向的意思。比如，假如我有了这个模型，我可以知道在当下的环境下，下一步的环境状态和reward是什么。这样，我就不必去真实的走这一步，就已经知道结果了，也就是不用非得试错了。这是个新的发展方向。

关于evolution method 要多说点：它和强化学习的区别在于，它不利用任何你的过程信息，只使用结果。比如我采用某一个policy,我就用这固定的policy和环境进行多次实验，看看最后的结果概率分布，然后知道这个policy有多大概率赢。然后换下一个，继续大量实验。最后在policy空间里找到一个最优的。它的缺点是忽略了大量的实验过程信息，也即根本没有考虑到value function。

２、tik-tok-toc游戏中的实例

关于value function的更新规则，"temporal difference learning method"

$$V(s) \leftarrow V(s) + \alpha \left[ V(s') - V(s) \right]$$

if the step-size parameter is reduced properly over time, this method converges, for any fixed opponent, to the true probabilities of winning from each state given optimal play by our player

从这个游戏中，可以引申出几个点的思考：

（１）先验知识的运用，可能改善学习效果

（２）强化学习的动作，除了像这个游戏这种离散的，也可能是连续的，reward函数也可能是连续函数。

（３）强化学习的状态集可能比这个游戏所有的大的多，如何保证在大的状态集上表现良好（具备很强的泛化能力），监督学习是一个好途径。

（４）如果能获得或者学习到一个环境模型，那么会更好的改善学习效果。