基于值函数的深度强化学习算法研究进展

摘　要 强化学习是当前机器学习的一个比较新的领域。跟监督学习和无监督学习相比，强化学习不是通过大量的静态数据学习一种映射关系，而是通过不断与环境进行交互，不断“试错”，动态的学习环境状态到所采取策略的映射关系。本文首先从强化学习的基本概率出发，主要以基于值函数的强化学习算法主线，阐述深度强化学习的研究发展状况，包括各种深度强化学习算法的演进，以及应用前景，最后简要介绍其发展方向。

关键词 强化学习，Value-base，Q-learning

第一章 引言

受行为驱动学科的启发，强化学习是关于智能体与环境交互的一种学习方式。环境处于一定的状态下，智能体通过一种行为与环境产生互动，然后环境给与智能体奖励，如何最大化这种奖励就是强化学习要解决的问题。20世纪末，伴随着机器人应用的研究发展，强化学习也逐渐称为热门的研究领域。特别是近十几年来随着深度学习的研究，越来越多的研究与深度学习相结合取得了突破性的进展，如[hoho：Hinton的Reducing the dimensionality of data with neural networks]用RBM深度神经网络实现对图像的降维，取得比PCA[hoho: PCA引用]更好的效果，被誉为首次深度学习兴起的成功实践。强化学习也不例外，通过与深度学习的结合，解决了很多复杂的连续状态空间、连续动作空间的问题，为机器人、金融、生物工程等领域开创了新局面。

强化学习可以分为基于模型的学习（model-base）和无模型的学习（model-free）两大类。基于模型的算法需要预先知道环境的状态转移函数和奖励函数，或者可以根据智能体与环境的交互采样数据学习到，譬如用动态规划的策略迭代和价值迭代算法，经典的Dyna-Q算法[hoho: 引用]等。无模型的算法则相反，不知道环境的模型参数，而是直接通过智能体与环境的交互数据直接学习策略或状态价值，如DQN、策略梯度、DDPG、PPO、SAC[hoho: 引用]等，都是这类型的算法。两种学习方法各有优缺点。通常在确定性环境中会使用基于模型的学习算法，如某些具有严格规则的棋牌类游戏。但这种白盒环境在现实中很少，很难对这种复杂环境建立良好的模型，这时使用无模型的学习算法会更加容易训练。 本文介绍的基于值函数的算法也都是属于无模型类型的学习算法。

本文只关注基于值函数的强化学习算法，是由于这类算法让人更容易进入强化学习这个领域，它更直接，更具代表性。以下从这几方面阐述基于值函数深度强化学习算法的研究进展：

- 第二章回顾强化学习的基本概念；

- 第三章列举近几十年来基于值的深度强化学习算法进展，并阐明算法的基本流程；
- 第四、五章阐述强化学习的应用领域与发展前景，总结当前强化学习领域需要解决的问题。

第二章 强化学习基本概念回顾

在介绍深度强化学习研究进展之前，我们先回顾一下强化学习的基本概念。

1. 马尔可夫决策过程（以下简称为MDP）

MDP是强化学习的基础理论。一个MDP描述为：在环境状态为s_t时，智能体采取动作a_t与环境进行交互，环境反馈智能体奖励r_t，并转移为新的状态s_{t+1}，如图1所示。

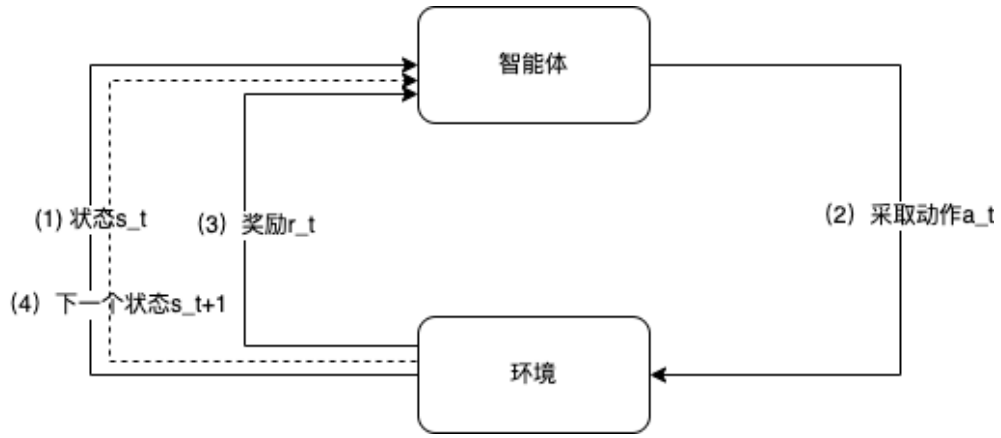

图1 一次马尔科夫决策过程

所以，MDP可描述为一个五元组：

$\{S, P, A, r, \gamma\}$，其中：

- S：状态的集合，$s_t \in S$；
- P：状态转移概率，如$p(s_{t+1}|s_t, a_t)$表示在状态为s_t（时间步t的状态）采取动作a_t的条件下，下一个时间步的状态为s_{t+1}的概率；
- A：动作的集合，$a_t \in A$；
- $r(s)$：某状态s下的奖励函数
- γ：折扣因子，$\tau \in [0, 1]$

2. 策略π

策略表示智能体在状态s下所采取动作a的概率分布，一般表示为$\pi(a_t|s_t) = p(a_t|s_t)$。策略也分为：

- 确定性策略：每个状态下只能输出一个动作，只有该动作概率为1，其他动作为0
- 随机策略：每个动作都服从一定的概率分布

3. 累积回报G

累积回报是智能体在每个状态下（一直到终结状态）获得的奖励的衰减之和：$G_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$，其中$\gamma$为折扣因子。折扣因子的存在是为了避免奖励之和的无穷大而使得智能体缺乏探索的动力。

4. 价值函数

状态的期望回报称为这个状态的价值函数（简称为V函数）：$V_\pi(s_t) = \mathbb{E}_\pi[G_t|s_t]$，其中$\pi$是当下所采取的策略。将状态价值函数的公式展开，就可以得到状态价值函数的递归表达：

$$
\begin{aligned}
V_\pi(s_t) &= \mathbb{E}_\pi[G_t|s_t] \\
&= \mathbb{E}_\pi[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + ...|s_t] \\
&= \mathbb{E}_\pi[r_t + \gamma(r_{t+1} + \gamma r_{t+2} + ...)|s_t] \\
&= \mathbb{E}_\pi[r_t + \gamma G_{t+1}|s_t] \\
&= \mathbb{E}_\pi[r_t + \gamma V(s_{t+1})|s_t] \\
&= r(s_t) + \gamma \sum_{s_{t+1} \in S} p(s_{t+1}|s_t) V(s_{t+1})
\end{aligned}
$$

在特征状态下采取动作的的期望回报称为状态-动作价值函数（也简称为动作价值函数或Q函数）：$Q_\pi(s_t, a_t) = \mathbb{E}_\pi[G_t|s_t, a_t]$

两者之间的关系为：$V_\pi(s_t) = \mathbb{E}[Q(s_t, a_t)] = \sum_{a_t \in A} \pi(a_t|s_t) Q_\pi(s_t, a_t)$。

强化学习的目的就是要最大化状态价值函数或动作价值函数。

5. 贝尔曼方程

贝尔曼是一位美国数学家，他提出并验证了著名的贝尔曼方程，成为了强化学习的基础理论。可以通过推导得出关于两个价值函数的贝尔曼期望方程：

$$
\begin{aligned}
Q_\pi(s_t, a_t) &= \mathbb{E}[r_t + \gamma Q_\pi(s_{t+1}, a_{t+1}|s_t, a_t)] \\
&= r(s_t, a_t) + \gamma \sum_{s_{t+1} \in S} p(s_{t+1}|s_t, a_t) \sum_{a_{t+1} \in A} \pi(a_{t+1}|s_{t+1}) Q_\pi(s_{t+1}, a_{t+1})
\end{aligned}
$$

$$
\begin{aligned}
V_\pi(s_t) &= \mathbb{E}[r_t + \gamma V_\pi(s_{t+1})|s_t] \\
&= \sum_{a_t \in A} \pi(a_t|s_t)(r(s_t, a_t) + \gamma \sum_{s_{t+1} \in S} p(s_{t+1}|s_t, a_t) V_\pi(s_{t+1}))
\end{aligned}
$$

另外还有贝尔曼最优方程，可以求出最优的两个价值函数：

$$
V^*(s_t) = \max_{a_t \in A} \{r(s_t, a_t) + \gamma \sum_{s_{t+1} \in S} p(s_{t+1}|s_t, a_t) V^*(s_{t+1})\}
$$

$$
Q^*(s_t, a_t) = r(s_t, a_t) + \gamma \sum_{s_{t+1} \in S} p(s_{t+1}|s_t, a_t) \max_{a_{t+1} \in A} Q^*(s_{t+1}, a_{t+1})
$$

有了贝尔曼方程，我们就可以用动态规划的思想求解马尔科夫决策问题了。

第三章 基于值函数的强化学习算法

1. DQN
2. DDQN
3. Dueling DQN
4. DRQN
5. DRAQN
6. H-DQN
7. Actic-Critic框架

第四章 应用与前景

第五章 总结