平衡探索与利用的方法

2023年10月31日　20:28

UCB

UCB（Upper Confidence Bound，上置信界）算法它通过使用置信上限来进行最佳动作的选择。UCB算法的基本公式如下：

$\mathrm{UCB_i} = \bar{\mathrm{X}}_\mathrm{i} + \mathrm{C} \times \sqrt{\mathrm{lnt}/N_i}$

在这个公式中：

- $\bar{X}_i$是动作 i 的平均奖励值。
- C 是控制探索和利用之间平衡的参数。
- t 是当前总的模拟或时间步数。
- N_i是动作 i 被选择的次数。

UCB值主要包括两项，前者表示当前动作-收益的实际分布，也就是实际的Q函数，后者则是对该动作不确定的一种度量。UCB的目标则是最大化动作的置信度，也就是置信区间，即表示为最大化公式，后者中N_i表示动作 i 被选择的次数，lnt表示选择动作总次数的对数，c是一个权值。简单地说，前者代表着开发，后者代表着探索。当当前动作被采样的次数很低时，N_i不变，而 lnt 在增加，探索部分变大，使得其被选择的概率越大；反之亦然。

UCT

UCT（上限置信区间树）通过计算置信上限来选择最优的行动。UCT的公式如下：

$\mathrm{UCT_i} = \bar{\mathrm{X}}_\mathrm{j} + \mathrm{C} \times \sqrt{\mathrm{lnN}/N_j}$

在这个公式中：

- $\bar{X}_j$是节点 j 的平均奖励值。
- C 是一个控制探索和利用之间权衡的参数。
- N 是总的模拟次数。
- N_j是节点 j 被访问的次数。

这个公式的含义是，它结合了节点的平均奖励值和置信上限项。平均奖励值表示了已知信息的价值，而置信上限项则考虑了探索未知信息的重要性，具体原理与UCB算法类似。通过这种方式，UCT算法能够选择在当前信息下最有潜力的行动，同时保持一定程度的探索，以便获得更多信息。

ϵ - greedy 策略

智能体做决策时，有一很小的正数 ϵ 的概率随机选择未知的一个动作，剩下1 − ϵ 的概率选择已有动过中动作价值最大的动作。基本公式如下：

$$
\begin{aligned}
&A^* \leftarrow \arg\max_a Q(s,a) \\
&\text{For all } a \in \mathcal{A}(s): \\
&\quad \pi(a|s) \leftarrow \begin{cases} 1-\varepsilon+\varepsilon/|\mathcal{A}(s)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(s)| & \text{if } a \neq A^* \end{cases}
\end{aligned}
$$

在决策过程中，有 ϵ 概率选择非贪心的动作，即每个动作被选择的概率为 ϵ/|A|，其中|A| 表示动作数量，另外还有1 − ϵ 的概率选择一个贪心策略，因此这个贪心策略被选择的概率则为1−ϵ+ϵ/|A|。