强化学习报告

周家豪

强化学习基础 理论 强化学习的形式化 马尔科夫决策过程 基于模型的方法 策略该代方法 无模型的方法(*) 基于策略迭代 -SARSA

强化学习理论之无模型的方法

周家豪

前瞻跨媒体实验室 计算技术研究所 ICT 中国科学院大学

jackhzhou@hotmail.com

2019年4月14日

目录

强化学习报告

周家園

强化学习基础 理论 强化学习的形式化 马尔科夫决方法 第略迭代方法 值迭代方法 值迭代方法 在被型的方法(*) 基于策略迭代 -SARSA 基于值迭

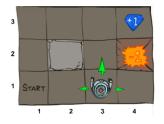
1 强化学习基础理论

- 强化学习的形式化
- 马尔科夫决策过程
- 基于模型的方法
 - 策略迭代方法
 - 值迭代方法
- 无模型的方法 (*)
 - 基于策略迭代 SARSA
 - 基于值迭代-Q-Learning
 - 总结

形式化-奖励

强化学习报告

選化学习基础 理论 學化學习的形式化 马尔科夫決策过程 基于模型的方法 值该代方法 无模型的方法 (*) 基于策略进代 -SARSA 基于强度进行。 E-Q-Learning 总结 "Reinforcement Learning focused on **goal-directed** learning from **interaction**."



Goal-directed - 奖励假设

"All goals can be described by the maximisation of expected cumulative reward."

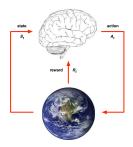
目标可以表示为奖励的累积。



形式化-状态, 行为

强化学习报告 _{周家豪}

"Reinforcement Learning focused on **goal-directed** learning from **interaction**."



Interaction - 交互建模

智能体看到环境**状态** S^t ,做出**行为** A^t ,得到 **奖励** R^t ,并到 新环境状态 S^{t+1} ,;

奖励度量智能体在状态下行为的好坏。

形式化-策略

强化学习报告

周家豪

强化学习基础 理论 强化学习的形式化 易尔科夫决策过程 基于模型的方法 策略迭代方法 值迭代方法 无模型的方法 等高选代方法 不规型的方法 等高选代方法 基于策略选代 -SARSA 基于链径 使-Q-Learning "Reinforcement Learning focused on **goal-directed** learning from **interaction**."



Solution - 求解目标

策略 $\pi(s) = a$ 是智能体在状态 s 选择行为 a 能得到最大累积 奖励 $\sum_t R^t$ 。

马尔科夫决策过程

强化学习报告

周家豪

强化学习的形式。 另外科夫决策过程 基于模型的方法 策略迭代方法 策略迭代方法 无模型的方法 传递代方法 无模型的方法 基于策略迭代 -SATE

Markov Decision Process

Markov Decision Process(MDP) is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$.

Markov Property

$$P(S^{t+1}|S^t, S^{t-1}, \dots) = P(S^{t+1}|S), S \in S$$

- S 环境的状态空间,
- **2** A 智能体的行为空间,
- 3 \mathcal{P} 状态转移函数 $\mathcal{P}_{ss'}^a = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a]$,
- 4 \mathcal{R} 奖励函数 $\mathcal{R}_s^a = \mathbb{E}[R_{t+1}|S_t = s, A_t = a],$
- 5 γ 折扣因子 $\gamma \in [0,1]$.

模型 - 马尔科夫决策过程的状态转移函数 \mathcal{P} 和奖励函数 \mathcal{R} 。

状态值函数

强化学习报告

周家

强化学习基础 设学习的形式过程 强化学习的形式过程 基于模型的方法 策略线代方法 无模型的声法 看选代方法(*) 基于模型的声法 基于模型的声法

最优策略 → 定义状态值函数评估最优策略。

状态值函数

$$v_{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t} \gamma^{t} R^{t} | S_{0} = s\right]$$

最优策略

策略偏序 - $\pi \ge \pi'$ if $\nu_{\pi}(s) \ge \nu'_{\pi}(s), \forall s$ 最优策略等价性 - $\nu_{\pi^*}(s) = \nu^*(s)$

策略迭代方法

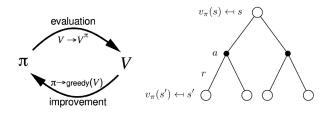
强化学习报告

周家豪

强化学习基础 理论

> 以外科夫决策过程 基于模型的方法 策略迭代方法 值迭代方法 后迭代方法 无模型的方法(*) 基于策略迭代 -SARSA

策略迭代 - 估计 + 控制 (提升) → 迭代策略求解。



$$egin{aligned} v_{\pi}(s) &= \mathbb{E}_{\pi}[R_{t+1} + \gamma v_{\pi}(S_{t+1}) | S_0 = s] ($$
贝尔曼期望方程 $) &= \mathcal{R}_s^a + \gamma \sum_{\sigma \in \mathcal{S}} \mathcal{P}_{ss'}^a V_{\pi}(s') \end{aligned}$

策略迭代算法

强化学习报告

周家豪

虽化学习基础 里论 强化学习的形式化 马尔科英族过程 基于模型的方法 情感代方法 传感代方法 无模型的方法 多不SARSA 基于價略迭代 -SARSA 《C-Q-Learning

```
Data: MDP\langle S, A, P, R, \gamma \rangle
Result: \pi^*(s) for all s
For each state s, initialize \pi(s) randomly;
repeat
      For each state s, initialize v_{\pi}(s) \leftarrow 0;
      /* 值估计
     repeat
           foreach s do
            v_{\pi}(s) \leftarrow \mathcal{R}_{s}^{\pi(s)} + \sum_{s'} \gamma \mathcal{P}_{ss'}^{\pi(s)} v_{\pi}(s')
           end
     until until v_{\pi}(s)'s converge;
      /* 策略提升 (控制)
                                                                                               */
     foreach s do
           \pi(s) \leftarrow \arg\max_{a \in \mathcal{A}} \mathcal{R}_s^a + \sum_{s'} \gamma \mathcal{P}_{ss'}^a v_{\pi}(s')
     end
until until \pi(s)'s converge;
```

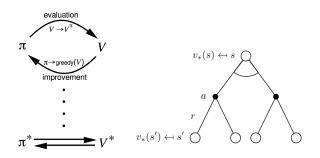
Algorithm 1: 策略迭代

值迭代方法

强化学习报告

周家豪

强化学习的影式化 身体等对的激式化 身体的表现的 身体的 基于模型的方法 **值迭代方法** 无模型的方法 无模型的方法 基于模型的方法 基于模型的方法 基于模型的方法 基于模型的方法 值迭代 - 最优值函数 (等价最优策略) → 迭代值函数求解。



$$v^*(s) = \max_a \mathcal{R}_s^a + \gamma \sum_{s' \in S} \mathcal{P}_{ss'}^a v^*(s')$$
(贝尔曼最优方程)

值迭代算法

强化学习报告

周家豪

```
贝尔曼最优方程 → 值函数迭代
```

Data: MDP $\langle S, A, P, R, \gamma \rangle$ **Result:** $\pi^*(s)$ for all s

For each state s, initialize $v^*(s) \leftarrow 0$;

repeat

$$v_{k+1}^*(s) = \max_{a \in \mathcal{A}} (\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v_k^*(s'))$$

end

until *until* $v^*(s)'s$ *converge*;

foreach s do

$$\pi^*(s) \leftarrow \max_{a \in \mathcal{A}} (\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a v^*(s'))$$

end

Algorithm 2: 值迭代

未知模型

强化学习报告

周家園

强化学习的形式。 理论 强化学习的形式。 马尔科夫决策过程 基于模型的方法 **值该代方法 在该代方法** 在该代方法 在该代方法 在该代方法 在该代方法 在该型的方法(*) 基于原础设代 - SAA 基于值选

未知模型

状态转移函数 $\mathcal{P}_{ss'}^a$ 计算复杂或者未知 奖励函数 \mathcal{R}_s^a 计算复杂或者未知 上述两种情况融合

状态转移函数 \mathcal{P} ,奖励函数 \mathcal{R} 未知 -> 能否估计 \mathcal{P},\mathcal{R} ? \to Model-based RL

基于模型的方法 Model-based RL

强化学习报告

周家

强化学习基础 理论 操化学习的形式化 马尔科夫决策过程 基于模型的方法 策略迭代方法 **值迭代方法** 无模型的方法(*) 基于策略迭代 -SARSA 基于值迭 代-Q-Learning

■ 抽样估计状态转移概率和奖励函数 → 例如: 蒙特卡洛估计 (Monte Carlo Estimation)

$$\hat{\mathcal{P}}_{ss'}^{a} = \frac{N_{ss'}^{a}}{N_{s}^{a}} \\ \hat{\mathcal{R}}_{s}^{a} = \frac{1}{M} \sum_{i=1}^{M} R_{s}^{a(i)}$$

2 利用估计的转移概率函数和奖励函数求解最优策略 能否直接估计值函数? → 无模型的方法。

策略迭代可行性

强化学习报告

周家豪

```
强化学习基础理论
强化学习的形式化
马尔科夫决策过程
基于模型的方法
策略该代方法
无模型的方法 (*)
基于策略选代-
SARSA
基于值选
代-Q-Learning
```

策略迭代方法在无模型是否可行?

- 1 估计 $v_{\pi}(s) \leftarrow \mathcal{R}_{s}^{\pi(s)} + \sum_{s'} \gamma \mathcal{P}_{ss'}^{\pi(s)} v_{\pi}(s')$ $v_{\pi}(s) = \mathbb{E}[\sum_{t} \gamma^{t} R^{t} | S^{0} = s; \pi]$ 可以估计 $v_{\pi}(s)$,例如蒙特卡洛估计 (MC)
- 2 控制 $\pi(s) \leftarrow \arg \max_{a \in \mathcal{A}} \mathcal{R}_s^a + \sum_{s'} \gamma \mathcal{P}_{ss'}^a v_{\pi}(s')$ 需要状态转移函数和奖励函数求解 \rightarrow 动作值函数

动作值函数

$$egin{aligned} q_\pi(s, \mathsf{a}) &= \mathcal{R}_s^{\mathsf{a}} + \sum_{\mathsf{s'}} \gamma \mathcal{P}_{\mathsf{ss'}}^{\mathsf{a}} v_\pi(\mathsf{s'}) \ q_\pi(s, \mathsf{a}) &= \mathcal{R}_s^{\mathsf{a}} + \sum_{\mathsf{s'}} \gamma \mathcal{P}_{\mathsf{ss'}}^{\mathsf{a}} q_\pi(\mathsf{s'}, \pi(\mathsf{s'}))$$
贝尔曼期望方程

策略迭代 (Q值)

```
Data: MDP\langle S, A, P, R, \gamma \rangle
强化学习报告
                     Result: \pi^*(s) for all s
                     For each state s, initialize \pi(s) randomly;
                     repeat
                           For each state s, initialize v_{\pi}(s) \leftarrow 0 Initialize
                            q_{\pi}(s,a)=0, \forall s,a
                           repeat
                                  foreach s and a do
 基于策略迭代 -
                                       v_{\pi}(s) \leftarrow \mathcal{R}_{s}^{\pi(s)} + \sum_{s'} \gamma \mathcal{P}_{ss'}^{\pi(s)} v_{\pi}(s')
                                      q_{\pi}(s,a) \leftarrow \mathcal{R}_s^a + \sum_{s'} \gamma P_{ss'}^a q_{\pi}(s',\pi(s'))
                                  end
                           until until \sqrt{\pi(s)}'s q_{\pi}(s, a)'s converge;
                           foreach s do
                                 \pi(s) \leftarrow \arg\max_{a \in \mathcal{A}} \mathcal{R}_s^a + \sum_{s'} \gamma \mathcal{P}_{ss'}^a v_{\pi}(s')
                                 \pi(s) \leftarrow \arg\max_{a} q_{\pi}(s, a)
                           end
                     until until \pi(s)'s converge;
```

Q 值估计

强化学习报告

周家園

```
强化学习基础
理论
强化学习的形式化
马尔特夫决策过程
基于模型的方法
值选代方法
后选代方法
无模型的方法(*)
基于等略选代 -
SARSA
基于值选
```

1 Monte Carlo Estimation

$$q_{\pi}(s,a) = \mathbb{E}[\sum_{t} \gamma^{t} R^{t} | S^{t} = s, A^{t} = a]$$
 $q_{\pi}(s,a) \leftarrow q_{\pi}(s,a) + \eta(\hat{q}_{\pi}(s,a) - q_{\pi}(s,a)), \hat{q}_{\pi}(s,a) = \frac{1}{M} \sum_{i=1}^{M} q_{\pi}(s,a)^{(i)}$ 样本更新 (低效)

2 Temporal Difference Estimation $q_{\pi}(s,a) = \mathbb{E}[R_t + \gamma q_{\pi}(s',a') | S^t = s, A^t = a]$ $q_{\pi}(s,a) \leftarrow q_{\pi}(s,a) + \eta((R_s^a + \gamma q_{\pi}(s',\pi(s'))) - q_{\pi}(s,a))$ 行

$$q_{\pi}(s,a) \leftarrow q_{\pi}(s,a) + \eta((R_s^a + \gamma q_{\pi}(s,\pi(s))) - q_{\pi}(s,a))$$
 为更新 (高效)

SARSA 算法 1

```
强化学习报告
```

周家豪

```
在化学习基础
理论

量化学习的形式化

局外有关决策过程

基于模型的方法

策酷迭代方法

统能代方法

统规则的方法(*)

基于策略迭代 -

SARSA

基于策略迭代 -

KARSA

基于策略迭代 -

KARSA
```

```
Data: \langle \mathcal{S}, \mathcal{A}, \gamma \rangle
```

Result: $\pi^*(s)$ for all s

For each state s and a, initialize $q_{\pi}(s, a)$ randomly;

foreach episode do

Set s to initial state; **repeat**

Take action $a \leftarrow \arg\max_{a'} q_{\pi}(s, a)$ Observe R, s' Choose action $a' \leftarrow \arg\max_{a'} q_{\pi}(s', a')$; $q_{\pi}(s, a) = q_{\pi}(s, a) + \eta[(R + \gamma q_{\pi}(s', \pi(s'))) - q_{\pi}(s, a)]$; $s \leftarrow s'$.

until until s is terminal state;

end

Algorithm 4: SARSA 算法 1

SARSA 收敛条件

强化学习报告

周家豪

```
强化学习的形式化
要化学习的形式化
曼化学习的形式化
身尔科夫决策过程
基于模型的方法
策略迭代方法
无
基于策略迭代 -
SARSA
基子值透
作、Q-Learning
```

收敛条件一 GLIE

 π is GLIE(Greedy in the limit with infinite exploration);

- I Greedy in the limit: 收敛到贪心策略; Take action $a \leftarrow \arg \max_{a} q_{\pi}(s, a)$ 满足;
- Infinite exploration: 所有 (s, a) 被抽样无穷次; Take action $a \leftarrow \arg \max_a q_{\pi}(s, a)$ 不满足;

状态空间不能穷尽,陷入局部最优 → 需要探索。

SARSA 收敛条件

强化学习报告

周家豪

```
强化学习基础
理论
强化学习的形式化
马尔科夫决策过程
基于模型的方法
策略或代方法
无模型的方法
(*)
基于策略选代 -
SARSA
基于值选
(*-Q-Learning
```

开发和探索

开发 - 根据已经探索的空间 (抽样数据) 计算最优策略,例如:策略迭代;

探索 - 获取未曾探索的状态行为空间,例如: ϵ -Greedy;

 ϵ -Greedy 以 ϵ 的概率选择随机行为,以 $1-\epsilon$ 的概率选择贪

心行为 \rightarrow 保证所有 (s, a) 被抽样无穷次;

 ϵ 逐渐减小,例如 $rac{1}{k} \longrightarrow$ 保证最终收敛到贪心策略。

收敛条件二 - Robbins-Monro

Robbins-Monro algorithm is a methodology for solving a root(θ^*) finding problem, where the function $M(\theta) = y$ is represented as an expected value $\mathbb{E}[N(\theta) = M(\theta)]$. $\theta_{t+1} = \theta_t - n_t(N(\theta) - y)$, $\lim_{t\to\infty} \theta_t \to \theta^*$

需要
$$\sum_t \eta^t = \infty, \sum_t (\eta^t)^2 < \infty$$
, 例如: $\eta^t = (\frac{1}{t})^p, p > 1$

SARSA 算法 2

```
周家豪

虽化学习基础

里论

强化学习的形式化

身环科夫埃莱过程

基于模型的方法

策略迭代方法

值或代方法

信或代方法

"在于模型的方法"

SARSA

基于值定

基于使型的方法"
```

强化学习报告

```
Data: \langle \mathcal{S}, \mathcal{A}, \gamma \rangle
Result: \pi^*(s) for all s
For each state s and a, initialize q_{\pi}(s, a) randomly;
foreach episode k do
     Set s to initial state:
     repeat
           Take action a \leftarrow \arg \max_a q_{\pi}(s, a) a by \epsilon-Greedy,
             Observe R, s' Choose action a' \leftarrow \arg \max_{a'} q_{\pi}(s', a') a'
            by \epsilon-Greedy:
          q_{\pi}(s, a) = q_{\pi}(s, a) + \eta[(R + \gamma q_{\pi}(s', \pi(s'))) - q_{\pi}(s, a)];
     until until s is terminal state;
end
```

Algorithm 5: SARSA 算法 2

值迭代方法可行性

强化学习报告

周家豪

```
强化学习基础
理论
强化学习的形式化
马尔科夫决策过程
基于模型的方法
策略迭代方法
值迭代方法
无模型的方法 (*)
基于策略迭代 -
SARSA
```

基于值迭

代-Q-Learning

值迭代方法在无模型时是否可行? v 值迭代

- 1 迭代 $v^*(s) = \max_{a \in \mathcal{A}} (\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a v^*(s'))$ 直到收敛; 难以直接估计 $v^*(s)$
- 2 求解 $\pi^*(s) \leftarrow \max_{a \in \mathcal{A}} (\mathcal{R}_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a v^*(s'))$ 需要模型 \mathcal{P}, \mathcal{R} 求解。

Q值迭代

- **1** 迭代 $Q^*(s,a) = R(s,a) + \sum_{s'} \mathcal{P}^a_{ss'} \max_{a'} \gamma Q^*(s',a')$ $q^*(s,a) \leftarrow q^*(s,a) + \eta[(R^a_s + \gamma \max_a q^*(s',a')) q^*(s,a)]$ 可以估计,例如 TD 估计
- 2 求解 $\pi^*(s) \leftarrow \arg \max_a Q^*(s, a)$ 无需模型求解。

Q-Learning 算法

```
强化学习报告
```

周家豪

```
異化学习基础

里轮

選化学习的形式化

場外科支架就理

基于模型的方法

策略迭代方法

信迭代方法

无模型的方法 (*)

基于策略迭代 -

SARSA

基于値迭

代・Q-Learning
```

```
Data: \langle \mathcal{S}, \mathcal{A}, \gamma \rangle
Result: \pi^*(s) for all s
For each state s and a, initialize q^*(s, a) randomly;
foreach episode do
     Set s to initial state:
     repeat
          Take action a by \epsilon-Greedy, Observe R, s' Choose action a'
           by \epsilon-GreedyChoose action a' by Greedy;
          q_{\pi}(s, a) = q_{\pi}(s, a) + \eta[(R + \gamma q_{\pi}(s', \pi(s'))) - q_{\pi}(s, a)];
          q^*(s, a) \leftarrow q^*(s, a) + \eta[(R_s^a + \gamma \max_{a} q^*(s', a')) - q^*(s, a)];
     until until s is terminal state;
end
```

Algorithm 6: Q-Learning 算法

SARSA VS Q-Learning

强化学习报告

虽化学习基础 理论 写尔科夫决策过程 基于模型的方法 第略迭代方法 伍选代方法 无模型的方法(*) 基于策略选代 -SARSA 基于镜的

总结

- 1 SARSA 源于策略迭代算法在无模型上的扩展。基于策略 迭代的 SARSA 的两次行为 (a, a') 来自于同样的策略, 定义为在线策略学习。
 - 若扩展 SARSA,使两次行为 (a, a') 来源于不同的策略,则为离线策略学习 (以不同的分布更新期望,需要进行重要性采样 (在不同分布下估计期望的方法))。
- 2 Q-Learning 源于值迭代算法在无模型上的扩展。基于值 迭代的 Q-Learning 的两次行为 (a, a') 来自不同的策略 $(\epsilon$ -Greedy 和 Greedy 策略),是离线策略学习。 Q-Learning 虽然是离线策略学习学习,但不需要重要性 采样即可保证其收敛性 (探索策略 + 学习率的选定)。

能否直接估计最优策略? → 策略梯度。

强化学习报告

周家家

强化学习基础 理论 强化学习的形式化 马尔科夫决策过的方法 策略迭代方法 值迭代方法 无模型的方法(*) 基于策略迭代 -SARSA 基于值选

总结

Thank You!