Paper 38. 透過反向強化學習進行學徒學習

38. Apprenticeship learning via inverse reinforcement learning

- **Introduction:**

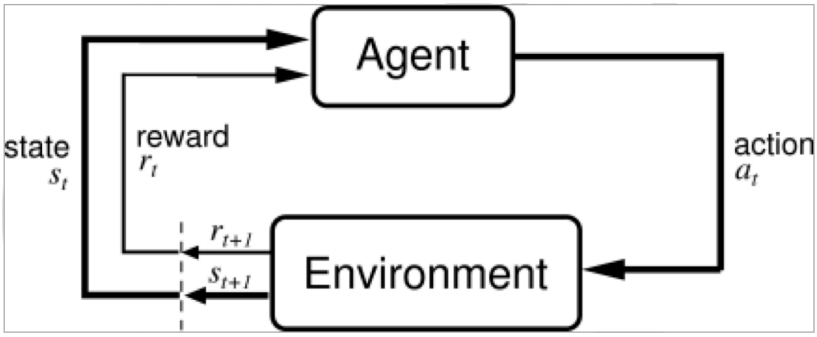
Agent從專家（expert）系統示範中找到或是學到非線性reward function 獎勵函數 R(s)=wT·φ(s)，使得Agent 找到或是學到的reward function 獎勵函數，其所學得的最佳化的policy 是跟專家（expert）示範的policy是幾乎是(非常)接近的。

Reward function獎勵函數R(s) 假設為

$$R(s)=w^T\cdot\varphi(s),$$

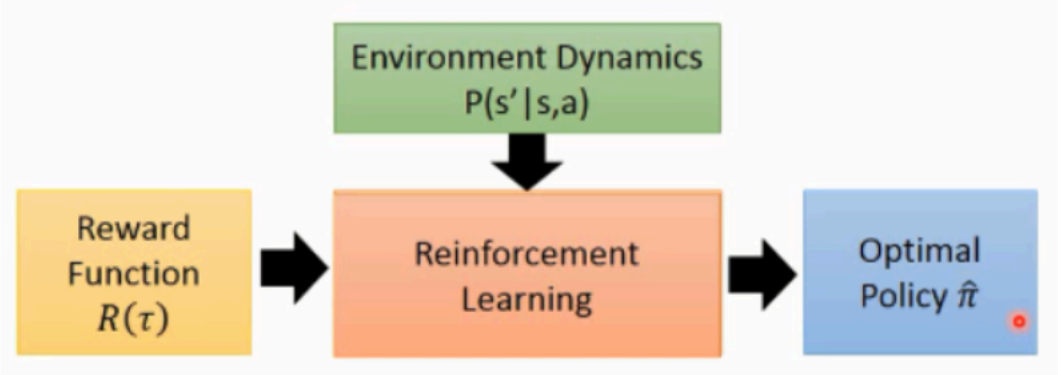
其中φ(s)為映射特徵的基函數,本論文是以線性函數為基底。

逆向強化學習求的就是reward function獎勵函數中的係數w。

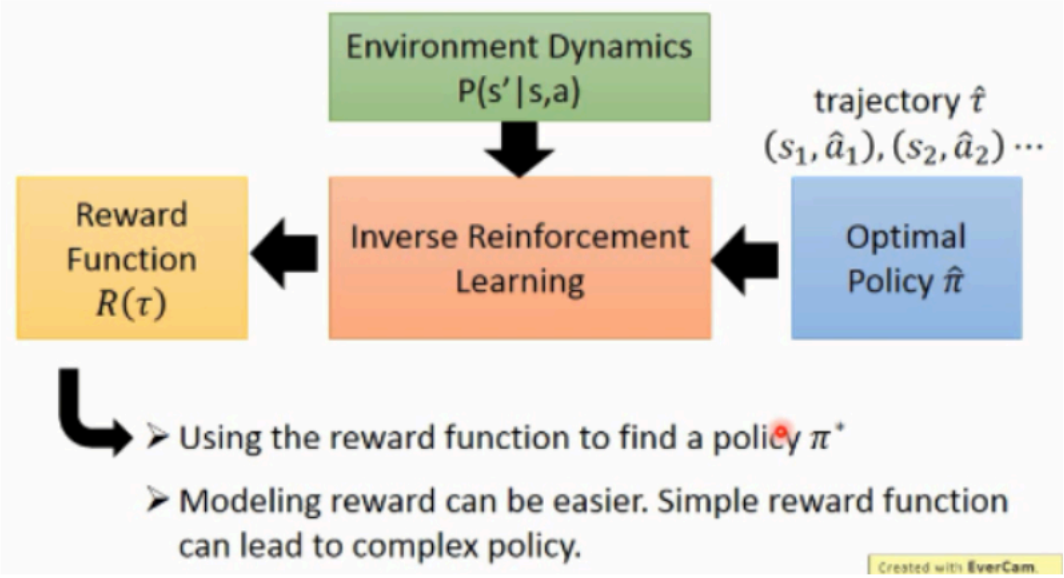


RL diagram

- RL: 一般RL是根據 reward 進行參數的調整，然後得到一個 policy



- IRL : IRL反向強化學習就不同了，因為他沒有顯示的 reward，只能根據人類行為，進行 reward的估計（反推 reward 的函數）


Environment Dynamics
P(s'|s,a)
trajectory $\hat{\tau}$
$(s_1, \hat{a}_1), (s_2, \hat{a}_2) \cdots$
Reward Function $R(\tau)$
Inverse Reinforcement Learning
Optimal Policy $\hat{\pi}$
Using the reward function to find a policy π^*
Modeling reward can be easier. Simple reward function can lead to complex policy.
Created with EverCam.

- **Problem Formula:**

策略Policy π 的value function為

$$v_\pi(s)= E_\pi[\sum_{t=0}^{\infty} \gamma^t R(s_t)]$$

將reward function代入：

$$v_\pi(s)= W^T E_\pi[\sum_{t=0}^{\infty} \gamma^t \phi(s_t)]$$

- 將上式右半部分定義為特徵期望值：$\mu(\pi)= E_\pi[\sum_{t=0}^{\infty} \gamma^t \phi(s_t)]$。需要注意的是，特徵期望跟策略 π 有關，策略不同時，策略期望也不相同

- 當給定m條專家軌跡後，根據定義我們可以估計專家策略的特徵期望為：

$$\hat{\mu}_E=1/m * \sum_{t=1}^{m} \sum_{t=0}^{\infty} \gamma^t \phi (s_t^{(i)})$$

其中，專家狀態序列為專家軌跡: $s_0^{(i)}, s_1^{(i)}, \cdots_{i=1}^{(m)}$

找到一個策略policy，使得該策略的表現與專家策略相近。我們可以利用特徵期望來表示一個策略的好壞，找到一個策略，使其表現與專家策略相近，其實就是找到一個策略$\tilde{\pi}$ 的特徵期望與專家策略的特徵期望相近，假若以下不等式成立：

$$\|\mu(\tilde{\pi})- \hat{\mu}_E \|_2 \leq \epsilon$$

當該不等式成立時，對於任意的權重$\|w\|_1 \leq 1$，Value Function滿足如下不等式：

$$| E [\sum_{t=0}^{\infty} \gamma^t R (s_t) |\pi_E] - E [\sum_{t=0}^{\infty} \gamma^t R (s_t) |\tilde{\pi}] | = | W^T\mu(\tilde{\pi}) - W^T\mu_E | \leq \|w\|_2 \|\mu(\tilde{\pi}) - \mu_E\|_2 \leq 1*\epsilon = \epsilon$$

最後希望減少policy去尋找讓$\mu(\tilde{\pi})$ 特徵期望能夠非常近似μ_E 的次數

- **Theoretical Analysis：**

從pseudo code來了解整個理論也許比較好理解

Pseudo code:

1. Randomly pick some policy $\pi^{(0)}$, compute or approximate via Monte Carlo $\mu^{(0)} = \mu(\pi^{(0)})$, and set $i = 1$.
2. Compte $t^{(i)} = \max_{w:\|w\|_2 \leq 1} \min_{j \in \{0,1,\ldots,m\}} w^T(\mu_E - \mu^{(j)})$ and let $w^{(i)}$ be the value of w that attains this maximum.
3. If $t^{(i)} \leq \epsilon$, then terminates.
4. Using RL algorithm, compute the optimal policy $\pi^{(i)}$ for the MDP using rewards $R = (w^{(i)})^T \phi(s)$.
5. Compute or estimate $\mu^{(i)} = \mu(\pi^{(i)})$.
6. Set $i = i + 1$ and go back to step 2.

其中step 2 的目標函數寫成標準的最佳化形式為

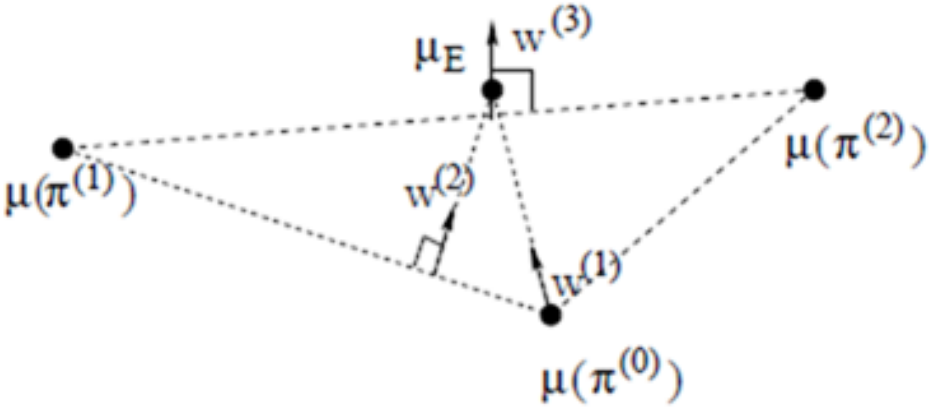
$$t^{(i)}=\max_{w:\|w\|_2 \leq 1} \min_{j\in 0\cdots(i-1)} w^T(\mu_E-\mu^{(j)})$$

寫成標準的最佳化形式為：

$$w^T \mu^{(j)} + t, j = 0, \ldots \leq \text{ s.t. } w^T\mu_E$$

$$\|w\|_2 \leq 1$$

在進行Step2求解時，$\mu^{(j)}$中的 $j \in 0,1,\cdots,i-1$是前i-1次迭代得到的最佳化策略。也就是說第i次求解參數時，i-1次迭代的策略是已知的。這時候的最佳化函數值t相當於專家策略 μ_E 與i-1個迭代策略之間的最大邊際。



我們可以從SVM(支持向量機)的角度去理解。專家策略為一類，其他策略為另一類，參數的求解其實就是找一條超曲面將專家策略和其他策略區分開來。這個超平面使得兩類之間的邊際最大。

Step4是在Step2求出參數後，便有了獎勵函數$R=(w^{(i)})^T\phi$，利用該獎勵函數進行強化學習，從而得到該獎勵函數下的最佳化的策略policy $\pi^{(i)}$。

最後可知反向強化學習學徒方法可分為兩步:

第一步在已經迭代得到的最佳化策略中，利用最大邊際方法求出當前的獎勵函數 R 的參數值；（該計算需要用到QP(二次規劃)機器或者 SVM機器。文中也給出了一種不使用SVM或QP機器的簡單算法。）

第二步利用求出的獎勵函數 R 的參數值進行正向強化學習方法求得當前最佳化的策略，然後重復第一步。

需要注意的是，$\phi(s)$ 中輸入的 s 為 I 個特徵：$s^1, s^2,\ldots, s^i$，
if
　　第 i 個特徵存在，$s^i=1$，
else
　　$s^i=0$。
又因為$\|w\|_1\leq 1$，所以$R\leq 1$

- **Conclusion：**

1. 技術上的限制

逆強化學習有可能會有policy overfitting 整個環境, 當環境不在是原先訓練的，整個Policy &Reward就需要再次重新迭代的訓練；缺點就是，如果想用到其他環境去，可能performance 會表現會不好。

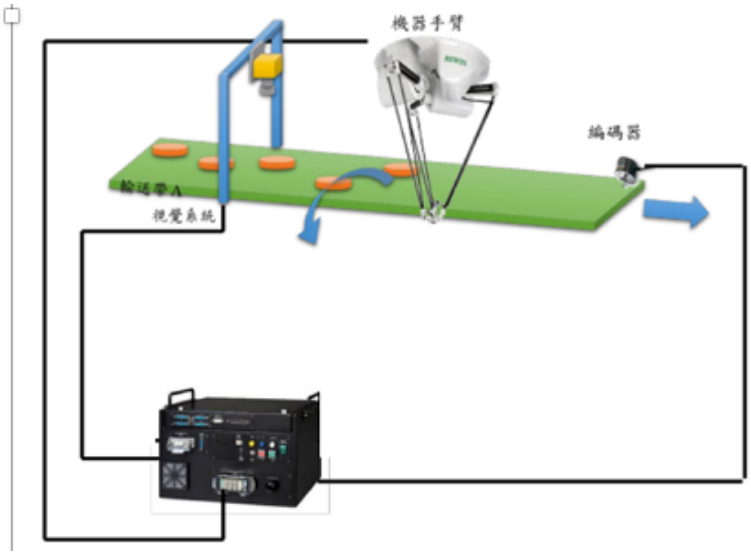
2. 潛在的未來研究方向

因為在工研院的計畫有準備要使用delta robot做PET寶特瓶回收(跟南部最大PET瓶回收商協治公司合作)，不知道是否逆強化學習（專家學習策略），是否可以能夠訓練出來一個很接近專家的手臂挑選系統（甚至比專家更厲害）.

最後，想針對複雜且多樣性的寶特瓶，使用RL進行挑選，並可以訓練出來比人更厲害的專家系統.

如下圖把這條輸送帶上的不是透明PET瓶子（環保署定義的第一類, 有bounding box就不是透明PET瓶）都用手臂挑走 .

YOLO v3 spp辨識的結果（有bounding box就不是透明PET瓶）



系統示意圖 :: 也許採用Hiwin(上銀) 1300 mm delta robot

- https://www.youtube.com/watch?v=0oXUoaN7WiY

- https://www.youtube.com/watch?v=Q7tE_vNYzzU

A 在A決定抓下一object 座標的策略(把不對PET[種類]的挑掉)

RL base (no model based, offline/online):

- Value-base: Q-learning (每一步都更新)
- Policy-base: Policy-Gradian (每回合更新)

RL base (model based, offline, include all no_model_base model)

Rule base:

- 最短路徑 B ,acc = 80, path 最短
- 兩者混合 D ,acc = 92, , path 中間
- 最高準確率. C ,acc = 99, path 可能最長 (現在正準備要做的)

ITRI Industrial Technology Research Institute

Robot 控制階段

- 第一階段 (~2020/8/31)
  - 只給A coordinator (ros topic: int16/ bounding box)
- 第二階段 (~2020/12/31)
  - 給A coordinator (ros topic: int16/ bounding box)
  - 給B,C,D (用rule base,最高Accuracy) coordinator
  - 給E,F,G,H coordinator
  - Call 機械所現有API ➔ A , BCD, EFGH
- 第三階段 (~2021/?/?)
  - 只給A,B,E 三個coordinator
  - 呼叫API , robot 8 parameter
    - 可控手臂4自由度 (前,後,左,右,上,下)
    - 控控加速度/煞車

- **Ref:**
  - Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. International Conference on Machine Learning(Vol.11, pp.1). ACM.
  - Format:

1 Introduction

Please provide a clear overview of the selected paper. You may want to discuss the following aspects:

- The main research challenges tackled by the paper
- The high-level technical insights into the problem of interest
- The main contributions of the paper (compared to the prior works)
- Your personal perspective on the proposed method

2 Problem Formulation

Please present the formulation in this section. You may want to cover the following aspects:

- Your notations (e.g. MDPs, value functions, function approximators,...etc)
- The optimization problem of interest
- The technical assumptions

3 Theoretical Analysis

Please present the theoretical analysis in this section. Moreover, please formally state the major theoretical results using theorem/proposition/corollary/lemma environments. Also, please clearly highlight your new proofs or extensions (if any).

4 Conclusion

Please provide succinct concluding remarks for your report. You may discuss the following aspects:

- The potential future research directions
- Any technical limitations
- Any latest results on the problem of interest

References