Abstraction Selection in Model-Based Reinforcement Learning

背景介绍

State abstraction的理论基础是06年的《Towards a Unified Theory of State Abstraction for MDPs》论文。

State abstraction是将高维的state压缩为低维的方法，假设已知一个有限的state abstraction集合，如果这个数据集很大，我们倾向于选择更加finer的abstraction（即更低的approximation error），因为其更忠于原模型；如果这个数据集很小，那我们倾向于选择更加coarser的abstraction（即更低的estimation error），因为其更能简化学习。

在这篇论文中，有一个前提：**假设数据集是已经确定了的。**由于本文只讨论abstraction的问题，因此在这里不考虑策略的影响，所以假设策略是等价最优的。

符号定义

MDP为$M=\langle S, A, P, R, \gamma\rangle$，$V$值为$V_M^*(s)=\max_{a \in A} Q_M^*(s, a)$，$Q$值为$Q_M^*(s, a)=R(s, a)+\gamma\left\langle P(s, a, \cdot), V_M^*(\cdot)\right\rangle$。

在数据集D和abstraction h 下建立的模型为M_D^h，x 为abstraction state，D 是四元组(s, a, r, s') 的集合。$M_D^h=\left\langle h(S), A, P_D^h, R_D^h, \gamma\right\rangle$

我们的目标是在候选集$\mathcal{H}$ 中选取一个合适的abstraction h 来最小化M_D^h 的loss：
$\operatorname{Loss}(h, D)=\left\|V_M^*-V_M^{\pi_M^*}\right\|_{\infty}$

损失函数的界限

在$M^h=\left\langle h(S), A, P^h, R^h, \gamma\right\rangle$ 中，

$$P^h\left(x, a, x'\right)=\frac{\sum_{s \in h^{-1}(x)} p(s, a) \sum_{s' \in h^{-1}\left(x'\right)} P\left(s, a, s'\right)}{\sum_{s \in h^{-1}(x)} p(s, a)}$$

$$R^h(x, a)=\frac{\sum_{s \in h^{-1}(x)} p(s, a) R(s, a)}{\sum_{s \in h^{-1}(x)} p(s, a)}$$

其中的transition error为：
$\epsilon_T^h=\max_{s \in S, a \in A} \sum_{x' \in h(S)}\left|P^h\left(h(s), a, x'\right)-\sum_{s' \in h^{-1}\left(x'\right)} P\left(s, a, s'\right)\right|$

reward error为：
$\epsilon_R^h=\max_{s \in S, a \in A}\left|R^h(h(s), a)-R(s, a)\right|$

对于任意h 而言，$\forall \delta \in (0,1)$,以$\geq 1-\delta$ 的概率有，

$$\mathrm{Loss}(h,D) \leq \frac{2}{(1-\gamma)^2}(\mathrm{Appr}(h)+\mathrm{Estm}(h,D,\delta))\ ,$$

$$\text{其中 } \mathrm{Appr}(h)=\epsilon_R^h+\frac{\gamma R_{\max}\epsilon_T^h}{2(1-\gamma)}$$

$$\mathrm{Estm}(h,D,\delta)=\frac{R_{\max}}{1-\gamma}\sqrt{\frac{1}{2n^h(D)}\log\frac{2|h(S)||A|}{\delta}}$$

$$n^h(D)=\min_{x\in h(S),a\in A}|D_{x,a}|$$

从这个式子中可以看出，$\mathrm{Appr}(h)$ 与$(\epsilon_T^h,\epsilon_R^h)$ 有关，而与数据集D 无关；$\mathrm{Estm}(h,D,\delta)$ 与ϵ_T^h 和ϵ_R^h 都无关，但是与$n^h(D)$ ——即the minimal number of visits to any abstract state-action pair——有关，也与$|h(S)|$有关。 因此当abstraction比较准确的时候ϵ_T^h和ϵ_R^h比较小，所以Appr(h)比较小；当abstraction比较概括的时候$h(S)$会比较小，因此Estm比较小。

总结

这篇文章就是在已知最优policy的情况下，通过该policy进行sample数据集，针对该数据集进行state abstraction。本文提出的方法是在一个$\mathcal{H}$的h候选集中选择一个最好的abstraction，并且理论证明了该h的误差上界。