# Towards a Unified Theory of State Abstraction for MDPs

强化学习有一个子领域是主要研究如何将高维的state表示进行降维操作的。在机器学习中，由于特征空间的维度往往很大，很难进行分类或回归学习。这时候，考虑到特征的一些冗余性和稀疏性，人们往往先对特征进行降维，然后在低维空间进行学习，这样做往往会取得好的效果。而state abstraction就是进行state的降维操作。这篇文章是2006年发表的，总结的是很传统的一些state abstraction(也叫state aggregation)方法，并提出了一个统一的理论和符号表示，分析了5种针对不同标准的策略各自的优劣以及state abstraction的一些性质，并给出了这些性质的证明，还算是值得看一下。

## 背景介绍

State abstraction 是从*ground* representation和original description映射到*abstract* representation中，即可以用来去除*irrelevant*的information。

MDP可以被定义为一个五元组$< S, A, P, R, \gamma >$。里面的各个符号的意思符合正常的MDP定义。$V^{\pi}(s)$ 和$Q^{\pi}(s,a)$ 也是正常公式中的意思。

带有*的即代表是optimal的。如$V^{*}(s) = \max_{\pi} V^{\pi}$和$Q^{*} = \max_{\pi} Q^{\pi}$

首先本文介绍了传统的很多种state聚类的方法：

| Abstraction Mechanism | Criterion | Exactness | MDP given?* | Notes |
|---|---|---|---|---|
| Bisimulation [11] | Model equivalence | Exact | Yes | Strictest measure |
| Homomorphisms [23] | Model equivalence | Exact, matching of actions flexible | Yes | Accounts for spacial relations (e.g. symmetry) |
| Approximate Bisimulation [5] | Model similarity | Bounded | Yes | Builds BMDPs |
| Bisimulation Metrics [8] | Model similarity | Statistically tested | Yes | Error bounds deducible |
| MAXQ [6] | Model equivalence for hierarchically consistent policies | Exact | Yes | Integrated into the MAXQ hierarchy |
| Stochastic Dynamic Programming [3] | Equivalent models given a policy | Exact | Yes | Covered by bisimulation |
| G Algorithm [4] | Equivalent rewards and Q-values | Statistically tested | No | Each feature's relevance must be independent |
| Utile Distinction [18] | Equivalent best actions with similar Q-values | Statistically tested | No | Aggregation occurs online |
| Policy Irrelevance [14] | Equivalent best actions | Statistically tested | No | May not yield optimal policy for ground MDP |
| Adaptive Aggregation [2] | Similar Bellman residuals | Bounded | Yes | States can be (dis)aggregated dynamically |

[3]: 随机动态规划，在factored设置中构建aggregation trees，以创建一个abstraction模型，其中固定策略下具有相同转移函数和奖励函数的状态被组合在一起。

[11]: 对于MDP给出了一个flat表示，同样也可以将固定策略下具有相同转移函数和奖励函数的状态进行聚合。

[23]: 基于模型的homomorphism(同态)进行状态聚合，并且在[29]中利用了option作为封装抽象信息的方式。

[5]: 类似于[11]中的做法，但是这里不要求精确的等价性要求。

[8]: 提出了bisimulation矩阵，通过相似矩阵来衡量两个状态的转移函数的相似性。将转移函数的相似性与奖励函数的相似性相结合可以对状态进行聚合。

[2]: 在基于Bellman残差的值迭代的基础上进行状态的分类。

[6]: 当奖励和转移函数在层次一致的任何策略中都相同时才有效。

[18]: 根据统计检验，只聚合具有相同最优action和类似Q值的状态。这种方法允许在系统学习MDP本身时进行这种聚合。

[4]: 和[18]类似。

[14]: 基于"Policy Irrelevance"对state分类

## 理论结构

### 定义

一般的MDP定义为$< S, A, P, R, \gamma >$，而abstract的MDP定义为$\langle \bar{S}, A, \bar{P}, \bar{R}, \gamma \rangle$。其中$\phi$ 为$S$ 到$\bar{S}$ 的映射，是一个可逆函数。这里作者说为了保证$\bar{P}$ 和$\bar{R}$ 能够被正确的定义，因此需要定义个权重函数$w$，其中$\sum_{s \in \phi^{-1}(\bar{s})} w(s) = 1$。遵从该定义，我们可以得到$\overline{P}$ 和$\overline{R}$：

$$\bar{R}^a_{\bar{s}} = \sum_{s \in \phi^{-1}(\bar{s})} w(s) R^a_s$$

$$\bar{P}^a_{\bar{s}\bar{s}'} = \sum_{s \in \phi^{-1}(\bar{s}) s' \in \phi^{-1}(\bar{s})} w(s) P^a_{ss'}$$

作者提到，之所以要定义权重函数$w$，实际上是因为我们要保证$\sum_{\bar{s}'} \bar{P}^a_{ss'} = 1$， 那么在这里

$$\begin{aligned}\sum_{\bar{s}'} \bar{P}^a_{ss'} &= \sum_{\bar{s}'} \sum_{s \in \phi^{-1}(\bar{s})} \sum_{s' \in \phi^{-1}(\bar{s})} w(s) P^a_{ss'} \\ &= \sum_{s \in \phi^{-1}(\bar{s})} w(s) \sum_{s' \in S} P^a_{ss'} \\ &= \sum_{s \in \phi^{-1}(\bar{s})} w(s) = 1\end{aligned}$$

所以这样定义$w$是正确的。即，假设$s_1$ 和$s_2$ 是属于同一个$\bar{s}$ 的。那么
$Q(\bar{s}, a) = w(s_1) \cdot Q(s_1, a) + w(s_2) \cdot Q(s_2, a)$。$w(s_1) + w(s_2) = 1$。

$w(s)$ 衡量了状态$S$ 对abstract的状态$\phi(s)$ 做的贡献。(可不可以考虑对于state进行attention计算？这里的$w(s)$就是state的attention系数。或者类似于prioritised ReplayMemory中的重要性之类的) 虽然本文里没有重点说$w$ 有什么用，但是$w$ 的重要性在[30]中有所体现。

对于policy而言：$\pi(s,a)=\bar{\pi}(\phi(s),a)$

## abstraction space的拓扑结构

- 在这里有两个拓扑学术语 **finer** 和**coarser** 。就是简单对于同一拓扑空间两个不同的拓扑的划分“细度”即包含关系的比较，当然前提是他们能有这样的包含关系，我们说他们是**可比较**的。

假设$\Phi_M$ 代表在MDP $M$ 中的abstraction空间，我们遵从如下定义：假设$\phi_1,\phi_2\in\Phi_M$ ，当对于所有的$s_1,s_2\in S$时，都有$\phi_1\left(s_1\right)=\phi_1\left(s_2\right)$ 能代表$\phi_2\left(s_1\right)=\phi_2\left(s_2\right)$ ，我们说$\phi_1$ finer than $\phi_2$ ，记为$\phi_1\succeq\phi_2$ 。反之则为coarser than。如果是strictly的就是去掉等号。

通过这些，我们可以得到$\phi_0$ 即原始MDP是finest的，其他的都要coarser than这个MDP。

## 5种聚类方式

1. 模型不变性的abstraction，即$\phi_{\text{ model}}\left(s_1\right)=\phi_{\text{ model}}\left(s_2\right)$ 代表了$R_{s_1}^a=R_{s_2}^a$ 和 $\sum_{s'\in\phi_{\text{model}}^{-1}\left(\bar{s}\right)}P_{s_1s'}^a=\sum_{s'\in\phi_{\text{model}}^{-1}\left(\bar{s}\right)}P_{s_2s'}^a$ 。直观上，该方式保留了one-step model（如[11]）。
2. $Q^\pi$ 不变性的abstraction，即$\phi_{Q^\pi}\left(s_1\right)=\phi_{Q^\pi}\left(s_2\right)$ 代表$Q^\pi\left(s_1,a\right)=Q^\pi\left(s_2,a\right)$。直观上，该方式保留了所有策略的state-action的value function。
3. $Q^*$ 不变性的abstraction，即$\phi_{Q^*}\left(s_1\right)=\phi_{Q^*}\left(s_2\right)$ 代表$Q^*\left(s_1,a\right)=Q^*\left(s_2,a\right)$ 。直观上，该方式保留了最优的state-action的value function（如[3]和[4]）。
4. $a^*$ 不变性的abstraction， 对于一个abstract类的state而言，都有同一个最优的$a^*$ ，并且 $\phi_{a^*}\left(s_1\right)=\phi_{a^*}\left(s_2\right)$ 代表着 $Q^*\left(s_1,a^*\right)=\max_aQ^*\left(s_1,a\right)=\max_aQ^*\left(s_2,a\right)=Q^*\left(s_2,a^*\right)$ 。直观上，该方式保留了最优的action和其value（如[18]）。
5. $\pi^*$ 不变性的abstraction， 对于一个abstract类的state而言，都有同一个最优的 $a^*$ ， $\phi_{\pi^*}\left(s_1\right)=\phi_{\pi^*}\left(s_2\right)$ 代表$Q^*\left(s_1,a^*\right)=\max_aQ^*\left(s_1,a\right)$ 和 $Q^*\left(s_2,a^*\right)=\max_aQ^*\left(s_2,a\right)$ 。直观上，该方式仅保留了最优action（如[14]）。

从论文中，没有太明确体现出4和5有什么不同之处。

## 性质

- 对于任意MDP，我们有$\phi_0\succeq\phi_{\text{ model}}\succeq\phi_{Q^\pi}\succeq\phi_{Q^*}\succeq\phi_{a^*}\succeq\phi_{\pi^*}$ 。即如果我们得到了一些关于$\phi_{\pi^*}$ 的性质，那也可以应用于所有比他finer的abstraction上面。
- 前4种abstraction（不算$\phi_0$）的最优policy都能够是原始MDP的最优policy。但是$\phi_{\pi^*}$ 的可能是原始MDP的次优policy。不过policy-search的方法可以避免这个问题。然而如果behavior policy不是fixed的，那么可能$\phi_{a^*}$ 和$\phi_{\pi^*}$ 都难以收敛。

针对第二点，作者进行了一定程度的讨论。他认为$\phi_{\pi^*}$ 的问题不是因为Q-learning是off-policy的，这种震荡问题可能是因为function approximator引入了partial observability导致的。

这里作者举了个例子，在$\phi_{a^*}$ 中，假设$Q(\bar{s}, a) = w(s_1) \cdot Q(s_1, a) + w(s_2) \cdot Q(s_2, a)$，由于这些权重$w$是基于state的visited频率计算的，所以其不是来自于一个平稳的分布，这会造成Q-learning的不稳定。而$\phi_{\pi^*}$ 也是类似的。所以说这两种方式都没有办法保证Q-learning的收敛，除非是用的fixed policy。这个结论不光适用于Q-learning，同样适用于其他的基于trajectory sampling并结合了non-stationary policy的方法。

- 对于$\phi_{model}$ ,$\phi_{Q^\pi}$ , $\phi_{Q^*}$ , $\phi_{a^*}$ 而言，只要权重函数$w(s)$是确定的，那么从经验而来建立的模型就会是收敛到真实的abstract模型。更广泛的来说，model-based RL能够收敛到greedy policy针对于原始MDP是最优的abstract value function上。但是$\phi_{\pi^*}$ 却不能保证这一点。

## Case Studies

| domains | $\phi_0$ | $\phi_{\text{model}}$ | $\phi_{Q^*}$ | $\phi_{a^*}$ | $\phi_{\pi^*}$ |
|---|---|---|---|---|---|
| TAXI $(5 \times 5)$ | 500 | 500 | 489 | 381 | 6 |
| COFFEE (v1) | 400 | 296 | 256 | 124 | 7 |
| COFFEE (v2) | 400 | 296 | 256 | 132 | 7 |
| BITFLIP (10) | 1024 | 513 | 257 | 11 | 1 |

Table 2: Sizes of abstract state spaces.

## 总结

实际上总结而言，我们在进行state abstraction的时候要考虑到minimizing information loss和maximizing state space reduction之间的平衡性问题。