Proto-value Functions: A Laplacian Framework for Learning Representation and Control in Markov Decision Processes

Sridhar Mahadevan MAHADEVA@CS.UMASS.EDU
Department of Computer Science
University of Massachusetts
Amherst, MA 01003, USA

Mauro Maggioni MAURO.MAGGIONI@DUKE.EDU
Department of Mathematics and Computer Science
Duke University
Durham, NC 27708

【强化学习 67】Proto-value Function

张楚珩
清华大学 交叉信息院博士在读

原文传送门

Mahadevan, Sridhar, and Mauro Maggioni. "Proto-value functions: A Laplacian framework for learning representation and control in Markov decision processes." Journal of Machine Learning Research 8.Oct (2007): 2169-2231.

特色

本文介绍了一种探索MDP上状态之间图结构的方法，能够有效探索到状态之间的对称性和瓶颈等信息；同时介绍了proto-value function（PVF）可以用作价值函数的基函数；提出了representation policy iteration（RPI），一边通过采样学习状态的表示（PVF），一边学习在这种表示下的价值函数策略。

本文的分析主要是在离散状态空间下做的，最后的部分还讲了推广到一类可分解的很大的状态空间和连续状态空间的方法。这里只讲里面最核心的部分，即一个MDP如何对应一个图模型，并且该图模型上的一类算子（diffusion matrix）是如何和MDP上的算子（Bellman operator以及transition matrix）相联系的。

过程

1. Spectral analysis of transition matrix

考虑离散的状态空间，各个量都用向量矩阵表示形式，对于一个策略 π，它的状态价值函数可以写为

$$V^{\pi} = (I - \gamma P^{\pi})^{-1} R^{\pi} = \left(I + \gamma P^{\pi} + \gamma^2 (P^{\pi})^2 + \ldots\right) R^{\pi}. \quad (1)$$

其中 P^{π} 表示策略 π 下的一步状态转移矩阵，其维度为 $|\mathcal{S}| \times |\mathcal{S}|$，它包括环境的随机性和策略的随机性；$R^{\pi}$ 表示从每个状态出发按照策略 π 行动之后得到的即时奖励，一般情况下即时奖励只和状态有关，那么它与策略无关。

一个矩阵反复乘到向量上得到的效果是，该向量会朝着该矩阵最大特征值对应特征向量方向偏移。下面研究价值函数的上述表达形式和 P^π 的特征向量有什么关系，具体说来，用 P^π 的特征向量基来表示价值函数。

考虑特征值分解

$$P^\pi = \Phi^\pi \Lambda^\pi (\Phi^\pi)^T,$$

写成向量形式

$$P^\pi = \sum_{i=1}^{n} \lambda_i^\pi \phi_i^\pi (\phi_i^\pi)^T,$$

其中为 ϕ_i^π 标准正交基。即时奖励也在该标准正交基下做分解，

$$R^\pi = \Phi^\pi \alpha^\pi, \tag{2}$$

价值函数可以用该标准正交基表示出来

$$\begin{aligned} V^\pi &= \sum_{i=0}^{\infty} (\gamma P^\pi)^i \Phi^\pi \alpha^\pi \\ &= \sum_{i=0}^{\infty} \sum_{k=1}^{n} \gamma^i (P^\pi)^i \phi_k^\pi \alpha_k^\pi \\ &= \sum_{k=1}^{n} \sum_{i=0}^{\infty} \gamma^i (\lambda_k^\pi)^i \phi_k^\pi \alpha_k^\pi \\ &= \sum_{k=1}^{n} \frac{1}{1-\gamma\lambda_k^\pi} \phi_k^\pi \alpha_k^\pi \\ &= \sum_{k=1}^{n} \beta_k \phi_k^\pi, \end{aligned}$$

注意到 P^π 的所有特征值的绝对值都小于1，因此，可以看出特征值最大的那一部分基底更为“重要”。由此，考虑使用 P^π 特征值最大的若干个基底来近似表示价值函数，即

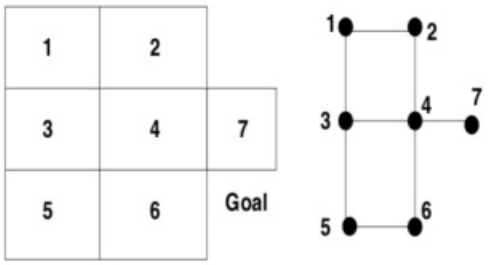
$$V^\pi \approx \sum_{k=1}^{m} \frac{1}{1-\gamma\lambda_k^\pi} \phi_k^\pi \alpha_k^\pi, \tag{3}$$

通过这种方式可以找到一系列的基底，使得价值函数能够被比较好地近似，下一步可以基于采集到的样本通过回归来找到相应的m个系数，由此估计得到价值函数。不过该方法存在几个问题：1）只有对称矩阵才能保证有实特征值（见【数值分析 1】矩阵分解），因此一个任意的 P^π 矩阵不一

定能做特征值分解；2）通常 P^{π} 是不知道的，求解它不比求解价值函数更简单。

2. Random walk

下面考虑下解决刚刚提到的第二个问题，基于采样来得到一个类似 P^{π} 的矩阵，然后对其做分解。建立一个邻接矩阵 W ，只要观察到两个状态之间能够一步转移，就把它们之间的连边记为 1 ，否则记为 0 。有了这样的图之后，构建一个随机游走（random walk）的一步转移矩阵 P_r ，容易知道 $P_r = D^{-1}W$ ，其中 D 是对角矩阵，其对角元为 W 中对应行的元素和。下面的例子可以帮助理解随机游走的一步转移矩阵 P_r 。



$$P_r = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0.33 & 0 & 0 & 0.33 & 0.33 & 0 & 0 \\ & & & \cdots & & & \end{bmatrix}$$

Figure 7: Top: A simple diffusion model given by an undirected unweighted graph connecting each state to neighbors that are reachable using a single (reversible) action. Bottom: first three rows of the random walk matrix $P_r = D^{-1}W$. P_r is not symmetric, but it has real eigenvalues and eigenvectors since it is spectrally related to the normalized graph Laplacian.

算子 P_r 和 P^{π} 有一定的相似之处，如果它们反复作用于一个代表初始状态分布的向量，得到的结果都是相应过程下的稳态状态分布。对于具有这样性质的算子称其为 diffusion model。算子 P^{π} 反复作用得到的稳态状态分布是在环境随机性和策略随机性共同作用下，很长时间以后，每一个时间步上所处状态的分布（假设 MDP 是 non-terminating 的）。同样，算子 P_r 反复作用得到的稳态状态分布是随机游走得到的相应分布。文章后面说明这样的算子包含了 MDP 本身的结构信息。

从上面的例子也可以看到，矩阵 P_r 不一定对称，因此它不一定有实特征值，由此我们想找到一个 diffusion model 使得它不仅反映 MDP 本身的结构信息，也同时保证能做实特征值分解。

3. Graph Laplacian

定义combinatorial Laplacian算子 $L = D - W$ 。

定义normalized Laplacian算子

$$\mathcal{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}.$$

可以发现它和随机游走算子 P_r 有着类似的特征系统（特征值和特征向量），具体地，假如算子 P_r 具有一个特征值 λ 和对应的特征向量 v，那么它就具有特征值 $1-\lambda$，并且对应的特征向量为 $D^{1/2}v$。并且，它是对称的。

由此可见可以选取算子 $\mathcal{L}$ 最小的若干个特征值对应的特征向量来作为价值函数的基。

4. Diffusion model 具有的性质

上述定义的Laplacian算子能够反映环境的一些几何特性，比如对称性（symmetries）和瓶颈（bottlenecks）。

定义automorphism来反映对称性，

Given a graph $G = (V, E, W)$, an *automorphism* π of a graph is a bijection $\pi : V \to V$ that leaves the weight matrix invariant. In other words, $w(u, v) = w(\pi(u), \pi(v))$. An automorphism π can be also represented in matrix form by a permutation matrix Γ that commutes with the weight matrix:

$$\Gamma W = W\Gamma.$$

可以得到算子 L 也含有该对称性

$$L\Gamma x = \Gamma L x = \Gamma\lambda x = \lambda\Gamma x.$$

定义Cheeger常数来反映瓶颈，

$$h_G(S) = \min_S \frac{|E(S, \tilde{S})|}{\min(\text{vol } S, \text{vol } \tilde{S})}.$$

Here, S is a subset of vertices, $\tilde{S}$ is the complement of S, and $E(S, \tilde{S})$ denotes the set of all edges (u, v) such that $u \in S$ and $v \in \tilde{S}$. The volume of a subset S is defined as $\text{vol } S = \sum_{x \in S} d_X$. Consider the problem of finding a subset S of states such that the edge boundary ∂S contains as few edges as possible, where

$$\partial S = \{(u, v) \in E(G) : u \in S \text{ and } v \notin S\}.$$

The relation between ∂S and the Cheeger constant is given by

$$|\partial S| \geq h_G \text{ vol } S.$$

算子 $\mathcal{L}$ 和瓶颈有一定的关系

Theorem 1 *(Chung, 1997): Define λ_1 to be the first (non-zero) eigenvalue of the normalized graph Laplacian operator $\mathcal{L}$ on a graph G. Let h_G denote the Cheeger constant of G. Then, we have* $2h_G \geq \lambda_1 > \frac{h_G^2}{2}$.

5. 例子

举一个实际的例子，如图所示，考虑一个两个房间的grid world，目标是通过“上下左右”的移动，移动到目标状态G，其最优价值函数如右图所示（转了180度）。

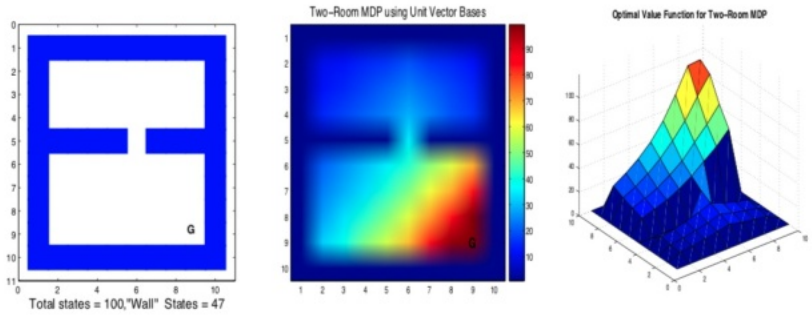
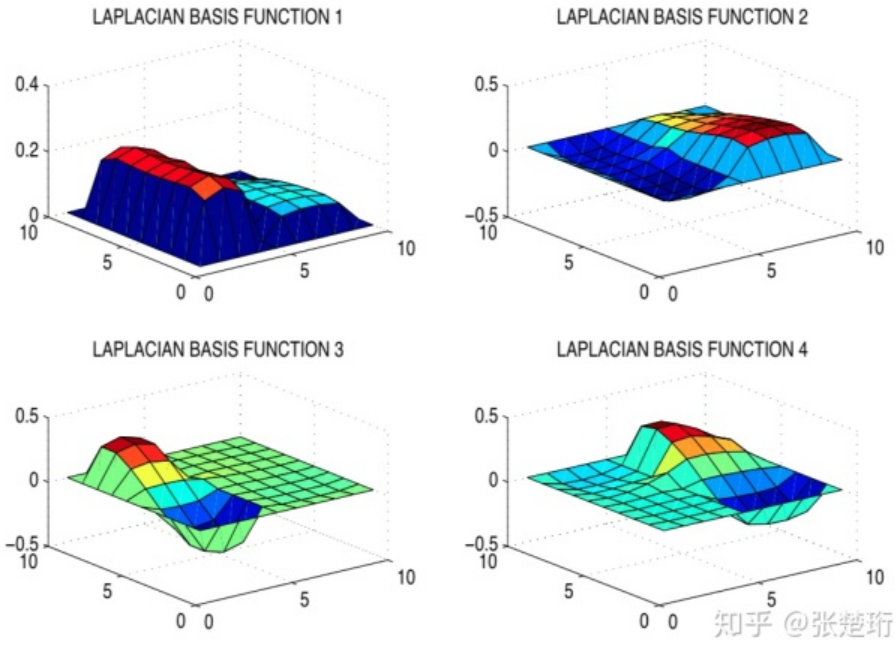


Figure 1: It is difficult to approximate nonlinear value functions using traditional parametric function approximators. Left: a “two-room” environment with 100 total states, divided into 57 accessible states (including one doorway state), and 43 inaccessible states representing exterior and interior walls (which are “one state” thick). Middle: a $2D$ view of the optimal value function for the two-room grid MDP, where the agent is (only) rewarded for reaching the state marked G by $+100$. Access to each room from the other is only available through a central door, and this “bottleneck” results in a strongly nonlinear optimal value function. Right: a $3D$ plot of the optimal value function, where the axes are reversed for clarity.

算子 $\mathcal{L}$ 的前几个基函数如下图所示（转了90度），比如，第一个基函数就区分开来两个房间，反映了状态的一定几何特性。



发布于 2019-06-07

强化学习 (Reinforcement Learning)

文章被以下专栏收录

强化学习前沿
读呀读paper