基于预测状态表示的 Q 学习算法

刘云龙，李人厚，刘建书
（西安交通大学系统工程研究所，710049，西安）

摘要：针对不确定环境的规划问题，提出了基于预测状态表示的 Q 学习算法. 将预测状态表示方法与 Q 学习算法结合，用预测状态表示的预测向量作为 Q 学习算法的状态表示，使得到的状态具有马尔可夫特性，满足强化学习任务的要求，进而用 Q 学习算法学习智能体的最优策略，可解决不确定环境下的规划问题. 仿真结果表明，在发现智能体的最优近似策略时，算法需要的学习周期数与假定环境状态已知情况下需要的学习周期数大致相同.

关键词：不确定环境规划；预测状态表示；Q 学习算法；奶酪迷宫
中图分类号：TP181 **文献标识码**：A **文章编号**：0253-987X(2008)12-1472-04

Q-Learning Algorithm Based on Predictive State Representations

LIU Yunlong, LI Renhou, LIU Jianshu
(System Engineering Institute, Xi′an Jiaotong University, Xi′an 710049, China)

Abstract: A Q-learning algorithm based on predictive state representations is proposed for solving the problem of planning under uncertainty. The predictive state representations is combined with the Q-learning algorithm. The prediction vector of predictive state representations is used as the state representation of Q-learning algorithms, so that the obtained states have the Markov properties and satisfy the requirement of reinforcement learning tasks. Then the Q-learning algorithm is used to find the optimal policy and the problem of planning under uncertainty is solved. Simulation results show that with our algorithm, the number of episodes needed in finding the near-optimal policy of an agent is approximately the same as that of the world states being assumed to be known.

Keywords: planning under uncertainty; predictive state representation; Q-learning algorithm; cheese maze

实际环境中的智能体，由于其感知能力有限，所以有可能感知不到所处环境的某些重要特征，同时智能体的动作也往往达不到预期的效果. 在这种局部可观测、随机的系统中，如何获取智能体的最优策略，即在不确定性环境下进行规划，是人工智能领域研究的一个重要问题.

局部可观测马尔可夫决策过程（POMDP）为解决不确定环境下的规划问题提供了丰富的数学框架[1]，但众所周知，POMDP 模型的建立，依赖于局部可观测的名义状态（Nominal State），所以学习系统的 POMDP 模型往往很困难，而且需要很多的先验知识.

预测状态表示法（PSRs）是最近提出的一种对受控动态系统建模的新方法[2]，与基于隐状态的 POMDP 模型不同的是，它能完全根据可观测的量来表现系统的状态，即利用系统中可执行的检验或实验所发生的概率组成向量，来表示系统的状态. 这样，通过观测数据，学习动态系统的预测状态表示（PSR）模型要比学习其 POMDP 模型更加容易，并且不容易陷于局部极小点[3].

收稿日期：2008-02-28. 作者简介：刘云龙（1977—），男，博士生；李人厚（联系人），男，教授，博士生导师. 基金项目：国家“211 工程”资助项目；教育部“985 工程”资助项目.

本文首先利用在人工智能和机器学习中通常碰到的奶酪迷宫问题具体地描述了预测状态表示方法的原理，然后以奶酪迷宫作为例子，在给定系统的 PSR 模型后，将 PSRs 与 Q 学习算法结合来解决不确定环境下的规划问题.

1　预测状态表示法

首先定义检验 $t=\{a^1o^1a^2o^2\cdots a^mo^m\}$ 在经历 $h=\{a^1o^1a^2o^2\cdots a^no^n\}$ 后发生的概率，即检验 t 在经历 h 后的预测值为 $\boldsymbol{p}(t|h)=\boldsymbol{p}(ht)/\boldsymbol{p}(h)=\text{prob}(o_{n+1}=o^1,o_{n+2}=o^2,\cdots,o_{n+m}=o^m\mid h,a_{n+1}=a^1,a_{n+2}=a^2,\cdots,a_{n+m}=a^m)$[3]. 一个经历指的是从初始时刻开始的一个动作-观测对序列；一个检验指的是一个动作-观测对序列，但不受从初始时刻开始的限定. 给定一系列检验集合 $Q=\{q_1,\cdots,q_k\}$，如果由这些检验的预测值所组成的向量

$$\boldsymbol{p}(Q\mid h)=[p(q_1\mid h),p(q_2\mid h),\cdots,p(q_k\mid h)]^{\mathrm{T}}$$

是所有经历的充分统计量，即 $\boldsymbol{p}(Q|h)$ 包含了经历 h 所有和未来预测相关的信息，也就是说，存在函数 f_t，对于所有经历 h，使得任意检验 t 发生的概率为

$$\boldsymbol{p}(t\mid h)=f_t(\boldsymbol{p}(Q\mid h))$$

则认为检验集合 Q 构成一个 PSRs. 其中，Q 称为检验核，$\boldsymbol{p}(Q|h)$ 称为预测向量，而 $\boldsymbol{p}(Q|h)$ 作为 PSRs 的状态表示. 可见，预测向量具有马尔可夫性. 同时，如果 f_t 是线性函数，则称该 PSRs 是线性 PSRs；如果 f_t 是非线性函数，则称该 PSRs 是非线性 PSRs. 本文针对的是线性 PSRs.

文献[2]表明，如果已获得系统的检验核 $Q=\{q_1,\cdots,q_k\}$，则对任意检验 t，存在长度为 k 的权向量 $\boldsymbol{m}_t$，使得对于任意经历 h，$\boldsymbol{p}(t|h)=\boldsymbol{p}^{\mathrm{T}}(Q|h)\boldsymbol{m}_t$. 该式表明，在得到经历为 h 的预测向量 $\boldsymbol{p}(Q|h)$ 之后，如果采取任意动作 a，得到任意观测值 o，则其对应的预测向量，即当前时刻的 PSR 状态的状态表示，可通过下式进行计算或更新，即对 $\forall q_i\in Q$[2]

$$\boldsymbol{p}(q_i\mid hao)=\frac{\boldsymbol{p}(aoq_i\mid h)}{\boldsymbol{p}(ao\mid h)}=\frac{\boldsymbol{p}^{\mathrm{T}}(Q\mid h)\boldsymbol{m}_{aoq_i}}{\boldsymbol{p}^{\mathrm{T}}(Q\mid h)\boldsymbol{m}_{ao}}\tag{1}$$

式中：$\boldsymbol{m}_{aoq_i}$ 是检验 aoq_i 的权向量；$\boldsymbol{m}_{ao}$ 是检验 ao 的权向量. 所以，经历 hao 的预测向量为[2]

$$\boldsymbol{p}(Q\mid hao)=\left(\frac{\boldsymbol{p}^{\mathrm{T}}(Q\mid h)\boldsymbol{M}_{ao}}{\boldsymbol{p}^{\mathrm{T}}(Q\mid h)\boldsymbol{m}_{ao}}\right)^{\mathrm{T}}\tag{2}$$

式中：$\boldsymbol{M}_{ao}$ 是一个 $k\times k$ 矩阵，第 i 列对应的是 $\boldsymbol{m}_{aoq_i}$. 根据式(2)可知，在得到 PSR 模型参数 $\boldsymbol{M}_{ao}$、$\boldsymbol{m}_{ao}$ 和空经历∅的预测向量 $\boldsymbol{p}(Q|\varnothing)$ 后，通过计算可得到任意经历的预测向量，即可得到任意经历下的状态表示. 根据以上描述可知，如果不同经历对应的预测向量相同，则不同经历对应同一个状态(PSR 状态). 如果不同经历对应的预测向量不同，则表示不同经历对应的 PSR 状态也不同.

在数学上，可用系统动态矩阵 $\boldsymbol{Z}$(见图 1)描述受控和非受控系统，而 PSR 模型可以直接由系统动态矩阵 $\boldsymbol{Z}$ 推导出来[3]. $\boldsymbol{Z}$ 的行对应所有可能的经历(过去)，其中第 1 行对应空经历，即初始经历；$\boldsymbol{Z}$ 的列对应所有可能的检验(未来)；$\boldsymbol{Z}$ 的元素表示在给定经历情况下检验发生的概率. 如果系统的动态矩阵的秩为 k，则矩阵中存在 k 个线性无关的检验列，则定义其为系统的检验核 Q. 同样，可将矩阵 $\boldsymbol{Z}$ 的行向量中任意一个最大线性无关组所对应的经历的集合，定义为经历核. 根据定义可知，检验核、经历核有可能不是惟一的.

	t_1	…	t_j	…
$h_1=\varnothing$	$p(t_1\|h_1)$	…	$p(t_j\|h_1)$	…
h_2	⋮			
h_i	$p(t_1\|h_i)$	…	$p(t_j\|h_i)$	…
⋮	⋮			

图 1　系统动态矩阵 $\boldsymbol{Z}$

2　举　例

下面以奶酪迷宫为例，描述预测状态表示方法的原理.

奶酪迷宫(见图 2)问题是人工智能和机器学习常常碰到的一个问题[4]. 迷宫中的智能体可以辨别初始状态(S)和目标状态(G)，但对于其他位置，只能根据位置四周是否有障碍物来确定. 图 1 中，数字表示智能体在迷宫中不同位置下得到的观测值，阴影部分为障碍物. 可以看出，有多个不同的环境状态，其所呈现的观测值相同. 系统的观测值集合为{1,2,3,4,5,6,7}，智能体的动作集合为{e(东)，s(南)，w(西)，n(北)}. 智能体在执行所选择的某一个动作时，如果在该动作方向上没有障碍物，智能体则沿动作方向移动一格；如果有障碍物，智能体则保持在原地不动.

获取系统的 PSR 模型往往需要先获取系统的检验核及经历核. 到目前为止，已有很多相关文献讨论了如何发现系统的检验核、经历核[5-6]. 由于本文的目的是描述 PSRs 的原理及其应用，所以假定当前系统的检验核、经历核已知. 奶酪迷宫系统的一个经历核为 $H=\{h_1=\varnothing,\ h_2=\text{w1},h_3=\text{e3},h_4=\text{e3e2},$

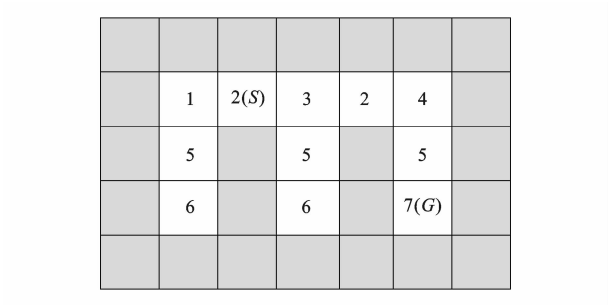

图 2 奶酪迷宫

h_5=e3e2e4，h_6=w1s5，h_7=e3s5，h_8=e3e2e4s5，h_9=w1s5s6，h_{10}=e3s5s6，h_{11}=e3e2e4s5s7}，一个检验核为 Q={q_1=w2，q_2=n1，q_3=n2，q_4=n3，q_5=n4，q_6=n5，q_7=s5，q_8=s6，q_9=e2，q_{10}=e3，q_{11}=n5n1}. 系统对应的检验核、经历核矩阵为

$$\boldsymbol{p}(Q\mid H)=\begin{bmatrix}0&0&1&0&0&0&0&0&0&1&0\\0&1&0&0&0&0&1&0&1&0&0\\1&0&0&1&0&0&1&0&1&0&0\\0&0&1&0&0&0&0&0&0&0&0\\1&0&0&0&1&0&1&0&0&0&0\\0&1&0&0&0&0&0&1&0&0&0\\0&0&0&1&0&0&0&1&0&0&0\\0&0&0&0&1&0&0&0&0&0&0\\0&0&0&0&0&1&0&1&0&0&1\\0&0&0&0&0&1&0&1&0&0&0\\0&0&0&0&0&1&0&0&0&0&0\end{bmatrix}$$

矩阵中的每个元素表示在当前经历下检验发生的概率，矩阵的行对应经历核中的经历，列对应检验核中的检验. 例如，第 1 行第 1 列表示空经历∅下检验 w2 发生的概率 $\boldsymbol{p}(q_1$=w2$\mid\varnothing)$=0，即在初始状态 S 为空经历∅下，采取动作 w 后，观测值 2 出现的概率为 0. 由矩阵 $\boldsymbol{p}(Q\mid H)$可知，初始预测向量为 $\boldsymbol{p}(Q\mid\varnothing)$=[0 0 1 0 0 0 0 0 0 1 0]$^{\mathrm{T}}$.

在得到系统的检验核、经历核及经历核-检验核矩阵 $\boldsymbol{p}(Q\mid H)$之后，通过以下公式可得到当前 PSR 模型的各个参数[4]

$$\begin{aligned}\boldsymbol{m}_{ao}&=\boldsymbol{p}^{-1}(Q\mid H)\boldsymbol{p}(ao\mid H)\\\boldsymbol{m}_{aoq_i}&=\boldsymbol{p}^{-1}(Q\mid H)\boldsymbol{p}(aoq_i\mid H)\end{aligned}\tag{3}$$

例如，参数 $\boldsymbol{m}_{s5}$可通过以下方式得到：由于 $\boldsymbol{p}$(s5$\mid H$)=[0 1 1 0 1 0 0 0 0 0 0]$^{\mathrm{T}}$，根据式(3)可得 $\boldsymbol{m}_{s5}=\boldsymbol{p}^{-1}(Q\mid H)\boldsymbol{p}$(s5$\mid H$)=[0 0 0 0 0 0 1 0 0 0 0]$^{\mathrm{T}}$.

同理，根据公式 $\boldsymbol{m}_{s5q_i}=\boldsymbol{p}^{-1}(Q\mid H)\boldsymbol{p}(s5q_i\mid H)$，可得 $\boldsymbol{m}_{s5q_i}(\forall q_i\in Q)$. 矩阵

$$\boldsymbol{M}_{s5}=\begin{bmatrix}0&-1&0&1&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0\\0&1&0&-1&1&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0\\0&0&0&1&-1&0&0&1&0&0&0\\0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0\end{bmatrix}$$

采用相同的方式，可计算得到系统的模型参数 $\boldsymbol{m}_{ao}$和 $\boldsymbol{M}_{ao}(\forall a\in A,o\in O)$，本文不再一一列出.

在得到系统的模型参数 $\boldsymbol{m}_{ao}$、$\boldsymbol{M}_{ao}(\forall a\in A,o\in O)$及初始预测向量 $\boldsymbol{p}(Q\mid\varnothing)$之后，系统在任意时刻对应的预测向量，即系统在任意时刻的预测状态表示，可根据式(2)计算得到. 例如

$$\begin{aligned}\boldsymbol{p}(Q\mid\text{e3s5})&=\left(\frac{\boldsymbol{p}^{\mathrm{T}}(Q\mid\text{e3})\boldsymbol{M}_{s5}}{\boldsymbol{p}^{\mathrm{T}}(Q\mid\text{e3})\boldsymbol{m}_{s5}}\right)^{\mathrm{T}}\\&=[0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0]^{\mathrm{T}}\end{aligned}$$

而 $\boldsymbol{p}(Q\mid$e3)可通过相同的方式计算得到，即

$$\boldsymbol{p}(Q\mid\text{e3})=\left(\frac{\boldsymbol{p}^{\mathrm{T}}(Q\mid\varnothing)\boldsymbol{M}_{e3}}{\boldsymbol{p}^{\mathrm{T}}(Q\mid\varnothing)\boldsymbol{m}_{e3}}\right)^{\mathrm{T}}$$

可见，在得到系统的初始预测向量及模型的各个参数后，任意经历 h 对应的预测向量均可通过式(2)逐项计算得到.

3 PSRs 与 Q 学习结合应用于不确定环境下的规划问题

如图 2 所示，实际环境中的智能体，由于其受到感知能力的限制，很可能感知不到所处环境的某些重要特征，这种现象称之为智能体遭遇隐状态[4]. 隐状态问题可能导致在不同的环境状态(World State)下得到同样的观测值，而不同的环境状态可能需要不同的动作，这种不确定环境下的规划问题是人工智能领域要解决的一个重要问题. 本文将 PSRs 与 Q 学习相结合，以解决不确定环境下的规划问题.

对于一个有限 PSR 状态系统，即当前系统可用有限阶的马尔可夫模型表示，都有一定数量的状态. 例如，第 2 节所讨论的奶酪迷宫系统，其 PSR 状态的数量为 11，即可能出现的不同预测向量的数量为 11. 在得到系统的 PSR 模型的各个参数后，如前所述，其预测向量具有马尔可夫特性，满足强化学习任

务的要求. 因此，可将预测向量作为 Q 学习算法的状态表示，并将 PSR 模型与 Q 学习算法结合，通过学习得到智能体的最优策略. 具体算法如下(对每一个学习周期重复以下步骤).

(1)将智能体置于初始状态 S，当前经历 h 为空经历$\varnothing$，其对应的预测向量为 $\boldsymbol{P}_{\text{Curr}}=\boldsymbol{p}(Q|\varnothing)$. 在当前预测向量下，根据每个动作所对应 $Q(\boldsymbol{P}_{\text{Curr}},a)$ $(\forall a\in A)$值的大小，选择动作 a. 其中，Q 学习算法中的动作选择采用ε-贪婪算法.

(2)在执行动作 a、得到观测值 o 之后，如果所处状态为目标状态，则一个学习周期结束. 否则，根据 PSR 模型计算当前经历 hao 所对应的预测向量[2]

$$\boldsymbol{P}_{\text{Next}}=\boldsymbol{p}(Q\mid hao)=\left(\frac{\boldsymbol{p}^{\mathrm{T}}(Q\mid h)\boldsymbol{M}_{ao}}{\boldsymbol{p}^{\mathrm{T}}(Q\mid h)\boldsymbol{m}_{ao}}\right)^{\mathrm{T}} \quad (4)$$

(3)$h\leftarrow hao$，并根据以下公式更新 $Q(\boldsymbol{P}_{\text{Curr}},a)$[7]，即

$$Q(\boldsymbol{P}_{\text{Curr}},a)\leftarrow Q(\boldsymbol{P}_{\text{Curr}},a)+\alpha(r+\gamma\max_{a\in A}Q(\boldsymbol{P}_{\text{Next}},a)-Q(\boldsymbol{P}_{\text{Curr}},a)) \quad (5)$$

式中：r 为当前得到的立即奖励值；α 是学习率；γ 是折扣因子. $\boldsymbol{P}_{\text{Curr}}\leftarrow\boldsymbol{P}_{\text{Next}}$.

(4)在当前预测向量下，根据每个动作所对应 $Q(\boldsymbol{P}_{\text{Curr}},a)$值的大小，采用$\varepsilon$-贪婪算法，选择动作 a，转入步骤(2).

一个学习周期指的是从初始状态到达目标状态. $Q(\boldsymbol{P}_{\text{Curr}},a)$是状态评估函数，指的是在预测向量 $\boldsymbol{P}_{\text{Curr}}$下执行动作 a，可获取的奖励值.

为了验证预测状态表示的有效性，对采用上述算法的智能体的学习效果与以下 2 种算法进行了比较、分析.

算法 1(理想情况)：假定环境状态已知，并以环境状态作为智能体所处状态，然后采用 Q 学习.

算法 2：以观测值作为智能体所处状态，然后采用 Q 学习.

为了增强算法的可比性，3 种算法采用 Q 学习的各个参数相同：学习率 $\alpha=0.125$，折扣因子 $\gamma=0.9$. 奶酪迷宫中的智能体要达到的目标是，从初始状态 S 开始以最短的路径到达目标状态 G. 对能使智能体到达目标状态的状态-动作给予的立即奖励值为 10，而对于其他的状态-动作不给予立即奖励.

针对 3 种算法，各进行了 10 次实验，每次的实验智能体执行 50 个学习周期，最后以 10 次实验得到的平均值作为各个学习周期内智能体到达目标状态所需要的步数. 图 3 表示理想情况下和采用预测状态表示作为状态表示的智能体的学习效果. 图 4 表示采用算法 2，即以观测值作为智能体所处状态的智能体的学习效果.

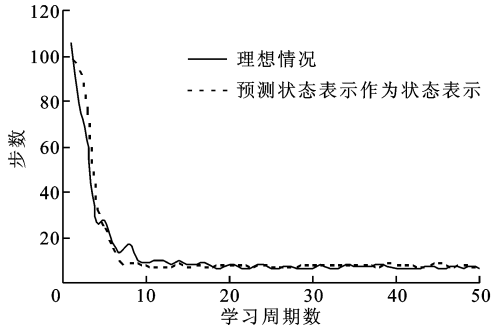

图 3　理想情况和以预测状态表示作为状态表示的智能体学习效果

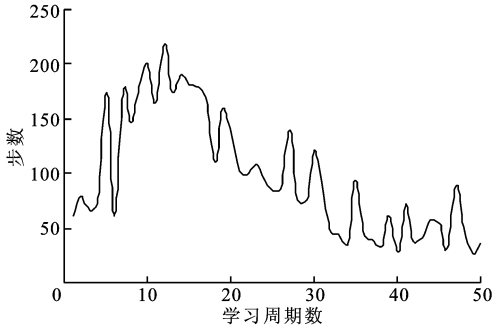

图 4　以观测值作为智能体所处状态的智能体学习效果

从图中可以看出，在发现智能体的最优近似策略时，基于预测状态表示的 Q 学习算法需要的学习周期数同假定环境状态已知，即理想情况下需要的学习周期数大致相同，而以观测值作为智能体所处状态的 Q 学习算法往往不能发现智能体的最优近似策略.

4　结　论

在对一个离散时间、有限观测值的受控动态系统建模时，PSRs 具有很多优点. 本文以奶酪迷宫问题为例，详细叙述了预测状态表示方法的原理. 然后，引入 Q 学习算法，以预测状态表示的预测向量作为 Q 学习算法的状态表示，将预测状态表示应用于不确定环境下的规划问题，并将其与采用不同状态表示的 Q 学习算法进行了比较分析. 仿真结果表明，采用预测状态表示的预测向量作为状态的 Q 学习算法，其学习效果与环境状态已知情况下的 Q 学习算法的学习效果基本吻合，并且远远优于直接以

观测值作为智能体状态的Q学习算法.同时,验证了基于预测状态表示的Q学习算法的有效性.

参考文献:

[1] KAELBLING L P, LITTMAN M L, CASSANDRA A R. Planning and acting in partially observable stochastic domains [J]. Artificial Intelligence, 1998, 101 (1/2): 99-134.

[2] LITTLEMAN M L, SUTTON R S, SINGH S. Predictive representation of state [M]// Advances in Neural Information Processing Systems 14. Cambridge, MA, USA: MIT Press, 2002: 1555-1561.

[3] SINGH S, JAMES M R, RUDARY M R. Predictive state representations: a new theory for modeling dynamical systems [C]// Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. Alberta, Canada: AUAI Press: 512-519.

[4] MCCALLUM A K. Reinforcement learning with selective perception and hidden state [D]. University of Rochester. Department of Computer Science, 1995.

[5] JAMES M R, SINGH S. Learning and discovery of predictive state representations in dynamical systems with reset [C]// Proceedings of the 21st International Conference on Machine Learning. New York, USA ACM, 2004: 417-424.

[6] WOLFE B, JAMES M R, SINGH S. Learning predictive state representations in dynamical systems without reset [C]// Proceeding of the 22nd International Conference on Machine Learning. New York, USA ACM, 2005: 985-992.

[7] SUTTON R S, BRATO A G. Reinforcement learning: an introduction [M]. Cambridge, MA, USA: MIT Press, 1998.

(编辑 苗凌)