文章编号: 1001-0920(2009)02-0289-04

基于 PSR 模型的规划算法

刘云龙, 李人厚, 刘建书
(西安交通大学 系统工程研究所, 西安 710049)

摘 要: 提出一种基于 PSR 模型的规划算法. 首先提出了状态经历的概念与发现方法, 并进一步用此概念来描述系统的 PSR 状态. 在此基础上, 讨论了如何用判别分析方法, 确定任意经历下的 PSR 状态以及如何在该过程中同时获取系统的 PSR 模型. 从而可引入 Q 学习算法, 用于决策当前的最优策略. 算法被应用于一些标准的 POMDP 问题, 实验结果验证了所提方法的有效性.

关键词: PSR 模型; 状态经历; 判别分析; Q 学习
中图分类号: TP181 **文献标识码**: A

Planning algorithm based on PSR models

LIU Yun-long, LI Ren-hou, LIU Jian-shu
(System Engineering Institute, Xi'an Jiaotong University, Xi'an 710049, China. Correspondent: LIU Yun-long, E-mail: ylliusv@163.com)

Abstract: A planning algorithm based on predictive state representation (PSR) models is proposed. The concept of state history is presented and used for describing the PSR state. Based on these, how to identify the PSR state at any history by using discriminant function analysis and how to obtain the PSR model of the system are discussed. Then, Q-learning algorithm is introduced for finding the optimal policy. The algorithm is applied to a standard set of POMDP test problems. Simulation results show the effectiveness of this algorithm.

Key words: PSR model; State history; Discriminant function analysis; Q-learning

1 引 言

预测状态表示(PSRs)是最近提出的一种对受控动态系统建模的方法[1]. 相比其他受控动态系统模型, PSR 模型具有很多优点[2]. PSRs 用系统中可执行的检验或实验所发生的概率组成向量, 以表示其状态. 如果存在一个检验, 它在两个经历下发生的概率不同, 则表示两个经历下系统的 PSR 状态不同; 如果对于所有可能发生的检验, 它们在两个经历下发生的概率均相同, 则表示两个经历下系统的 PSR 状态相同. 对于一个受控系统, 一个经历指的是从初始时刻开始的一个动作-观测值对序列. 同经历一样, 一个检验指的也是一个动作-观测值对序列, 但没有从初始时刻开始的限定. 已经证明, PSR 模型的状态表示, 不必用到所有检验发生的概率, 只用检验核(core-tests)中所有检验发生的概率即可.

基于 PSR 模型的规划是关于 PSRs 研究的一个很重要的方面, 同时也是研究 PSRs 的一个很重要的目的[3]. 对于一个未知系统, 用于计算其 PSR 模型的各个参数的数据, 是通过和系统交互而获取的, 因此得到的系统的 PSR 模型往往会存在噪声, 导致在不同经历下, 系统可能处于同一个 PSR 状态. 但是, 根据 PSR 模型计算得到的检验核中检验发生的概率却可能不同. 所以, 仅用计算得到的预测向量确定系统的 PSR 状态是靠不住的.

为解决上述问题, 本文首先提出了状态经历的概念及其发现方法. 在获得系统的 PSR 模型的同时, 用状态经历的集合来描述系统的所有 PSR 状态; 然后, 引入判别分析方法, 确定任意经历下的 PSR 状态; 再进一步, 通过与 Q 学习算法结合, 提出一种新的基于 PSR 模型的规划算法.

2 PSRs 简介

假定系统的观测值集合为 $O = \{o^1, o^2, \cdots, o^{|O|}\}$, 动作集合为 $A = \{a^1, a^2, \cdots, a^{|A|}\}$, 则定义长

收稿日期: 2007-12-05; **修回日期**: 2008-03-04.
基金项目: 国家"211 工程"项目; 西安交通大学"行动计划"项目.
作者简介: 刘云龙(1977—), 男, 山东安丘人, 博士生, 从事强化学习、环境建模的研究; 李人厚(1935—), 男, 浙江宁波人, 教授, 博士生导师, 从事智能控制理论与方法、环境建模等研究.

度为 m 的检验 $t = \{a^1o^1a^2o^2\cdots a^mo^m\}$ 在经历 $h = \{a^1o^1a^2o^2\cdots a^no^n\}$ 下发生的概率为

$$p(t\mid h) = p(ht)/p(h) = prob(o_{n+1} = o^1, o_{n+2} = o^2, \cdots, o_{n+m} = o^m \mid h, a_{n+1} = a^1, a_{n+2} = a^2, \cdots, a_{n+m} = a^m).$$

其中:a_i 表示在 i 时刻采取的动作;o_i 表示在 i 时刻执行动作 a_i 后出现的观测值. 给定一系列检验的集合 $Q = \{q_1, \cdots, q_k\}$,如果由这些检验的预测值所组成的向量 $p(Q\mid h) = [p(q_1\mid h), p(q_2\mid h), \cdots, p(q_k\mid h)]^{\mathrm{T}}$ 是所有经历的充分统计量,即 $p(Q\mid h)$ 包含了经历 h 下所有和未来预测相关的信息,也就是存在函数 f_t,对于所有经历 h,使得任意检验 t 发生的概率为 $p(t\mid h) = f_t(p(Q\mid h))$,则认为检验集合 Q 构成一个 PSRs. 其中 $p(q_i\mid h)$ 是在经历 h 下检验 q_i 发生的概率;Q 被称为检验核,$p(Q\mid h)$ 被称为预测向量,以 $p(Q\mid h)$ 作为 PSRs 的状态表示.

在数学上,可用系统动态矩阵 Z 描述受控和非受控系统,而 PSR 模型可以直接从系统动态矩阵 Z 推导出来[2]. Z 的元素表示在给定经历情况下,检验发生的概率. 如果系统的系统动态矩阵的秩为 k,则矩阵中存在 k 个线性无关的检验列,满足检验核的定义,可将其作为系统的检验核 Q. 同样,将矩阵 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称为经历核(core-histories). 如果已获得检验核,则对每一个检验 t,存在长度为 k 的权向量 m_t,使得相应于检验 t 的矩阵的列 $p(t\mid h)$,可表示为 $p(t\mid h) = p(Q\mid h)^{\mathrm{T}}m_t$. 该式表明,在得到经历为 h 的预测向量 $p(Q\mid h)$ 后,如采取任意动作 $a\in A$,得到任意观测值 $o\in O$,则其对应的预测向量,即当前时刻的状态表示可通过式(1)进行计算或更新. 即对 $\forall q_i\in Q$[1],有

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

其中:m_{aoq_i} 是检验 aoq_i 的权向量,m_{ao} 是检验 ao 的权向量. 所以经历 hao 的预测向量为[1]

$$p(Q\mid hao) = \frac{p(Q\mid h)^{\mathrm{T}}M_{ao}}{p(Q\mid h)^{\mathrm{T}}m_{ao}}, \tag{2}$$

其中 M_{ao} 是一个 $k\times k$ 矩阵,第 i 列对应的是 m_{aoq_i}. 由式(2)可知,在得到 PSR 模型参数 M_{ao},m_{ao} 和空经历 ϕ 的预测向量 $p(Q\mid\phi)$ 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.

3 基于 PSR 模型的规划

3.1 预备知识

定义 1 系统动态矩阵 Z 中,提取所有两两线性无关的预测向量行,它们所对应的经历所组成的集合称为 Z 的状态经历集合,集合中的每元素称为状态经历.

在可以精确获得系统中任意 $p(t\mid h)$ 的前提下,提出如下定理及推理:

定理 1 状态经历集合可以表示系统的所有 PSR 状态,并且状态经历集合中状态经历和系统的 PSR 状态集合中 PSR 状态是一一对应的.

证明 给定任意两个不同的 PSR 状态,h_1 和 h_2 分别为对应这两个 PSR 状态的经历. 假定这两个 PSR 状态的预测向量线性相关,则对任意检验 t,$p(t\mid h_1) = a * p(t\mid h_2)$,其中 a 为一个常数. 由于对两个不同的 PSR 状态,其对应的所有检验发生的概率不完全相同,可知 $a\neq 1$. 然而,根据 $\forall k\ \forall\bar{a}\in A^k$,$\sum_{t\in T(\bar{a})}p(t\mid h) = 1$(注:$T(\bar{a})$ 表示所有动作序列为 $\bar{a}$ 的检验集合,A^k 表示长度为 k 的所有动作的集合,h 为所有可能发生的经历)[2],可得

$$\sum_{t\in T(\bar{a})}p(t\mid h_1) = \sum_{t\in T(\bar{a})}a * p(t\mid h_2) = a * \sum_{t\in T(\bar{a})}p(t\mid h_2) = a * 1 = a.$$

因为 $\sum_{t\in T(\bar{a})}p(t\mid h_1) = 1$,得 $a = 1$,与前面 $a\neq 1$ 相矛盾,故任意两个不同 PSR 状态的预测向量是线性无关的. 根据状态经历的定义可知,状态经历集合可表示所有的 PSR 状态,并且不同的 PSR 状态对应不同的状态经历.

根据不同的状态经历对应的 PSR 状态不同,可知状态经历集合中状态经历和系统的 PSR 状态集合中 PSR 状态是一一对应的. □

根据定理 1 的证明可得如下推理:

推理 1 如果两个不同经历的预测向量线性相关,则两个经历表示同一个 PSR 状态.

3.2 任意经历下 PSR状态的确定及系统的 PSR模型的获取

本文采用蒙特卡罗方法获取 Z 的子矩阵中的元素[4],可以认为通过该方法得到的对应同一个 PSR 状态的不同经历 h 的预测向量构成了一个正态总体. 另外,通过估计得到的矩阵的元素存在噪声,因此本文按照文献[4]所提方法计算矩阵的秩.从而本文提出的任意经历下 PSR 状态的确定及系统的 PSR 模型的获取方法如下:

首先,定义系统动态矩阵 Z 的任意子矩阵 D 的经历扩展为:在第 1 步,得到矩阵 D 的状态经历集合 SH_1,其中状态经历的数量为 r_1;在第 $i(i\geq 2)$ 步,获取矩阵 D_i,其行对应的经历的集合为 $\{h, hao\mid \forall a, o, h\in SH_{i-1}\}$,其列为 D 对应的列 T_D. 然后得到 D_i 的状态经历集合 SH_i,其中状态经历的数量为

r_i. 如果 $r_i = r_{i-1}$,算法停止,并以 SH_i 作为矩阵 D 经历扩展后的状态经历集合 SH,得到状态经历 - 检验预测矩阵 $p(T_D \mid SH)$,然后计算其秩为 v 并获取当前发现的检验核 Q_T($p(T_D \mid SH)$ 中任意 v 个线性无关的列所对应的检验的集合). 其中矩阵的状态经历集合的获取方法为:根据推理 1 及表示同一个 PSR 状态的不同经历 h 的预测向量构成一个总体,将当前矩阵的所有行向量划分为若干组,其中每组中的行向量是两两线性相关的,而取自不同组的行向量是两两线性无关的. 将每组中的所有行向量标记为取自同一个总体的一个样本,对于每个样本中的所有经历,如果存在空经历,则将空经历标记为状态经历;否则,将发生次数最多的经历标记为状态经历.

然后,通过以下步骤发现系统的状态经历集合、经历核和检验核:

Step1:获取 Z 的子矩阵 Z_1,其行对应的经历的集合为 $H_1 = \{\phi\} \cup \{ao \mid \forall a, o\}$;其列对应的检验的集合为 $T_1 = \{ao \mid \forall a, o\}$. 执行矩阵 Z_1 的经历扩展,得到当前的状态经历 - 检验预测矩阵、检验核 Q_{T_1} 和状态经历 - 检验预测矩阵的秩 v_1.

Step i($i \geq 2$):重复以下过程:获取 Z 的子矩阵 Z_i,其行对应的经历的集合为 $H_i = H_1$;其列对应的检验的集合为 $T_i = \{ao, t, aot \mid \forall a, o, t \in Q_{T_{i-1}}\}$. 执行矩阵 Z_i 的经历扩展,得到当前状态经历 - 检验预测矩阵、检验核 Q_{T_i} 和状态经历 - 检验预测矩阵的秩 v_i. 如果 $v_i \neq v_{i-1}$,$i = i + 1$,继续执行 Step i;否则,算法停止,并以执行矩阵 Z_i 的经历扩展后得到的状态经历集合和检验核,分别作为发现的系统的状态经历集合 SH 和检验核 Q_T. 令 $k = v_i$,即 $\mathrm{rank}(p(Q_T \mid SH)) = k$.

Z_i 的经历扩展的最后一步,在获取扩展后的状态经历集合的过程中,每组两两线性相关的行向量已被标记为取自同一个总体的一个样本,而每个样本中有一个状态经历. 因此,如果 SH 中状态经历的数量为 N,则可得到 N 个取自不同总体的样本. 将所有样本中所有行向量的元素仅保留与检验核中检验相对应的部分. 如果 $N = k$,则可认为状态经历集合即经历核;如果 $N > k$,则对于矩阵 $p(Q_T \mid SH)$ 而言,存在 C_N^k 个 $k \times k$ 子矩阵. 计算每个子矩阵的条件数,将条件数最小的 $k \times k$ 子矩阵的行所对应的经历的集合作为经历核 Q_H.

最后,根据下式计算得到系统的 PSR 模型的各个参数[4]:

$$m_{ao} = p^{-1}(Q_T \mid Q_H)\, p(ao \mid Q_H),$$
$$m_{aoq_i} = p^{-1}(Q_T \mid Q_H)\, p(aoq_i \mid Q_H). \quad (3)$$

根据状态经历的获取原则,空经历(即初始状态的预测向量)会被保留.

同时,根据样本 i 估计总体 i 的均值向量 u_i 和协方差阵 $\Sigma_{(i)}$[5]. 此时,如果已知一个经历的预测向量,则其对应的 PSR 状态可通过判别分析方法[5] 获得,本文采用的是距离判别法.

3.3　基于 PSR 模型的使用 Q 学习的规划算法

首先,根据定理 1 可知状态经历的数量 N 即系统的 PSR 状态的数量,并且状态经历和 PSR 状态一一对应. 对所有动作 a,任意初始化 $Q(s_i, a)$,其中 s_i($i = 1, \cdots, N$) 为第 i 个 PSR 状态.

然后,对每一个学习周期重复以下步骤:

Step1　将智能体置于初始状态 s,$h \leftarrow \phi$. 在当前状态下,根据每个动作所对应 $Q(s, a)$($\forall a \in A$) 值的大小,采用 ε 贪婪算法,选择动作 a.

Step2　执行动作 a,得到观测值 o 后,如果所处状态为目标状态,则一个学习周期结束. 否则判断 hao 是否为状态经历,若是,则当前 PSR 状态为状态经历 hao 所对应的 PSR 状态,假定其为 s';若不是,则计算当前经历 hao 对应的预测向量[1]

$$p(Q \mid hao) = \frac{p(Q \mid h)^{\mathrm{T}} M_{ao}}{p(Q \mid h)^{\mathrm{T}} m_{ao}}. \quad (4)$$

然后采用距离判别法,计算当前智能体所处的 PSR 状态为

$$s' \leftarrow \arg\min_i [p(Q \mid hao) - u_i]^{\mathrm{T}} \Sigma_{(i)}^{-1} \times [p(Q \mid hao) - u_i],\ i = 1, \cdots, N. \quad (5)$$

Step3　$h \leftarrow hao$,更新 $Q(s, a)$[6],即

$$Q(s, a) \leftarrow Q(s, a) + \alpha(r + \gamma \max_{a \in A} Q(s', a) - Q(s, a)), \quad (6)$$

其中 r 为立即奖励值. 然后将 s' 对应的状态经历的预测向量 $p(Q \mid sh')$ 赋予经历 h,作为经历 h 的预测向量,$s \leftarrow s'$.

Step4　在 s 下,根据 $Q(s, a)$($\forall a \in A$) 值的大小,采用 ε 贪婪算法,选择动作 a,转入执行 Step2.

其中一个学习周期指的是从初始状态到达目标状态. $Q(s, a)$ 是状态评估函数,指的是在 PSR 状态 s 执行动作 a 可获取的奖励值. 式(6) 中的 α 是学习率,γ 是折扣因子.

4　实验研究

为了验证本文提出的算法的性能,将本文算法应用于一些标准的 POMDP 系统[7]. 下面叙述了实验内容和结果,并且将本文算法与利用 CMAC 对预测向量泛化后使用 Q 学习的规划算法[8] 做了比较分析.

针对每个系统,首先通过本文所提算法得到系

统的 PSR 模型以及状态经历的集合;然后,利用本文提出的基于 PSR 模型的 Q 算法学习每个系统中的最优控制策略. 在学习的过程中,每学习一定的步数,停止学习;然后利用学习到的策略在环境中采用 ε 贪婪算法执行 100 000 步,并以该过程中得到的累积的奖励值除以 100 000 作为对应学习步数的平均回报,并将其作为评价算法性能的一种指标[8]. 针对两种算法,各做了 10 次实验,试验结果如图 1 ~ 图 3 所示. 其中,横坐标表示学习的步数,纵坐标表示对应每个学习步数 10 次实验得到的平均回报的均值. 本文算法采用的 Q 学习的学习率 α = 0.125, 折扣因子 γ = 0.9, ε = 0.01. 对于利用 CMAC 对预测向量泛化后使用 Q 学习的规划算法,针对每个系统的 CMAC 的参数设置参见文献[8].

从图1～图3可以看出,对于3个不同的系统,本文算法均明显好于利用 CMAC 对预测向量泛化后使用 Q 学习的规划算法. 通过本文算法找到一个较好策略需要的步数,远远小于利用 CMAC 对预测向量泛化后使用 Q 学习的规划算法所需要的步数,验证了本文算法是有效的.

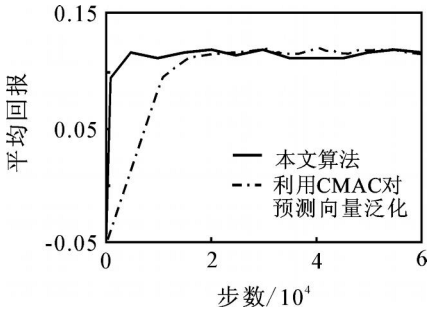

图 1 4 ×3 Maze 环境下获取的平均回报

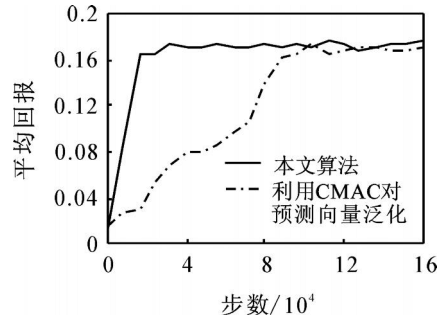

图 2 Cheese Maze 环境下获取的平均回报

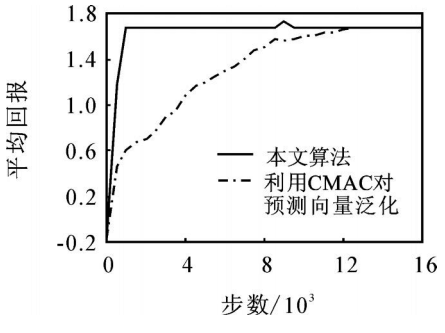

图 3 Shuttle 环境下获取的平均回报

5 结 论

基于 PSR 模型的规划是 PSRs 研究面临的重要问题. 本文提出了状态经历的概念,并证明了状态经历的集合可以描述系统的所有 PSR 状态,再借助判别分析方法,可确定任意经历下系统所处的状态. 通过它与强化学习算法的结合,提出了一种基于 PSR 模型的规划算法. 利用一些标准的 POMDP 系统,通过仿真结果表明了本文所提出的算法是有效的.

参考文献(References)

[1] Littleman M L, Sutton R S, Singh S. Predictive representation of state [C]. Advances in Neural Information Processing Systems 14. Vancouver: MIT Press, 2002: 1555-1561.

[2] Singh S, James M R, Rudary M R. Predictive state representations: A new theory for modeling dynamical systems[C]. Uncertainty in Artificial Intelligence: Proc of the Twentieth Conf. Banff: AUAI Press, 2004: 512-519.

[3] Rosencrantz M, Gordon G, Thrun S. Learning low dimensional predictive representations[C]. Proc of the Twenty-First Int Conf on Machine Learning. Banff, 2004: 695-702.

[4] James M R, Singh S. Learning and discovery of predictive state representations in dynamical systems with reset[C]. Proc of the Twenty-First Int Conf on Machine Learning. Banff, 2004: 417-424.

[5] 向东进, 李宏伟, 刘小雅. 实用多元统计分析[M]. 武汉: 中国地质大学出版社, 2005.
(Xiang D J, Li H W, Liu X Y. Applied multivariate statistical analysis[M]. Wuhan: CUGP Press, 2005.)

[6] Sutton R S, Brato A G. Reinforcement learning: An introduction[M]. Cambridge: MIT Press, 1998.

[7] Cassandra A. Tony, pomdp file repository page[EB/OL]. (2008-03-02). http://www.cs.brown.edu/research/ai/pomdp/examples/index.html.

[8] James M R, Singh S, Littleman M L. Planning with predictive state representations[C]. Proc of the Int Conf on Machine Learning and Applications. Louisville: IEEE Press, 2004: 304-311.