发现和学习不可复位动态系统的预测状态表示的一种新算法

刘云龙, 李人厚
(西安交通大学系统工程研究所, 陕西西安 710049)

摘 要: 提出了一种发现和学习不可复位动态系统的预测状态表示的新算法. 在证明系统的任意 landmark 均可作为系统的初始状态的基础上, 利用发现的 landmark 确定系统在任意时间步所处的经历, 然后采用蒙特卡罗方法估计任意经历下任意检验发生的概率, 解决了在不可复位动态系统中, 经历下检验发生的概率难以获取问题, 进而发现和学习不可复位动态系统的预测状态表示. 实验结果表明, 本文算法获得的系统的预测状态表示在预测精度上明显优于 suffix-history 算法, 验证了所提算法的有效性.

关键词: 预测状态表示; 不可复位动态系统; landmark; suffix-history 算法
中图分类号: TP181 文献标识码: A 文章编号: 0372-2112 (2009) 01-0126-06

A New Algorithm for Discovery and Learning of Predictive State Representations in Dynamical Systems Without Reset

LIU Yun-long, LI Ren-hou
(*System Engineering Institute, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China*)

Abstract: A new algorithm for discovery and learning of predictive state representations in dynamical systems without reset is proposed. With proving that any landmark can be used as the initial state, the discovered landmarks are used to identify the history at any time step in a continues data, then the conditional probability of any test at any history is estimated using Monte Carlo approaches, which efficiently solves the difficult problem of obtaining the conditional probability in dynamical systems without reset, thereby it is straightforward to discover and learn predictive state representations. The empirical results show that in case of the obtained predictive state representations' s prediction quality, our algorithm has better prediction accuracy than the suffix-history algorithm, which proves the effectiveness of the proposed algorithm.

Key words: predictive state representations; dynamical systems without reset; landmark; suffix-history algorithm

1 引言

对动态系统建模, 是科学和工程中普遍存在的一个问题, 在很多方面都有很广泛的应用. 通过对系统建模, 可以控制和预测该系统. 例如, 一个投资者可以利用一个公司以往的股价, 对该公司的股价建立一个模型, 从而预测该公司未来股价的走势, 决定未来投资的方向. 具体到人工智能领域, 对于一个随机的、局部可观测的环境, 智能体由于受到其感知能力的限制, 有可能会感知不到所处环境的某些重要特征, 同时, 智能体采取的动作也往往得不到预期的效果. 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——即不确定环境下的规划问题, 是人工智能领域研究的一个重要问题[1]. 常用的解决这种问题的途径是首先对系统建模, 然后根据得到的系统模型对问题进行求解.

局部可观测马尔可夫决策过程(Partially Observable Markov Decision Processes, POMDPs)为解决这种不确定性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), 因此学习系统的 POMDP 模型往往很困难, 并且需要很多先验知识[2].

预测状态表示(Predictive State Representations, PSRs)是最近提出的一种对受控动态系统建模的方法[3]. 和基于隐状态(hidden-state-based)的 POMDP 模型不同, PSRs 完全根据可观测的量来表现系统的状态. 通过观测数据学习动态系统的 PSR 模型比学习其 POMDP 模型应该会

收稿日期: 2007-12-11; 修回日期: 2008-06-04
基金项目: 国家"211 工程"资助; 西安交通大学"行动计划"资助

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

更容易, 且不易陷于局部极小点[2]. 同时, PSRs 克服了另一种常用的动态系统模型——基于经历的模型(history-based model) 存在的严重局限性, 即很多可用 PSR 模型表示的环境, 不能用有限阶次的基于经历的模型来表示[2].

PSRs 用可以在系统中执行的检验(tests) 或实验(experiments) 发生的概率所组成的向量表示其状态. 如果存在一个检验, 它在两个经历(history) 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-观测对序列. 同经历一样, 一个检验指的也是一个动作-观测对序列, 但没有从初始时刻开始的限定[3]. 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就行. 详细见第 2 节.

研究 PSRs, 需要解决以下三个主要问题: (1) 发现: 发现用来表示状态的检验的集合, 即发现检验核. (2) 学习: 学习系统的预测状态表示的各个参数, 即系统的 PSR 模型的各个参数. (3) 规划: 建立基于 PSR 模型的规划[4]. 受控动态系统可以分为可复位的系统和不可复位的系统. 本文主要研究不可复位动态系统的预测状态表示的发现和学习问题.

在现实环境中, 很多系统是不可复位的, 即不存在一个动作, 使得不论系统当前处于何种状态, 通过该动作后均可以使系统回到原来的初始状态或初始配置[5]. 因此, 用于发现和学习系统预测状态表示的训练数据, 只有一个连续的动作-观测值对序列, 而不像可复位系统中那样, 可以得到一组从初始状态开始的动作-观测对序列. 如何利用这一个连续的动作-观测对序列来发现和学习不可复位动态系统的预测状态表示, 是 PSRs 研究的一个重要问题. 现有的针对不可复位系统的PSR 模型的发现和学习方法主要有 suffix-history 算法[5,6], 该算法将一个训练数据中的每个时间步都看作是一个新训练样本的起点, 然后通过蒙特卡罗方法, 估计经历下检验发生的概率. 由于它将每个时间步都看作是一个新训练样本的起点, 有可能导致利用该算法确定的经历往往不是系统实际所处的经历, 从而使得得到的预测向量实际上是多个不同经历的预测向量的加权和. 并且, 只有在特定条件下, 通过该方法发现的 PSR 模型与系统本身的 PSR 模型一致. 同时, 该方法对策略的选择也有特定的要求.

本文提出了一种发现和学习不可复位动态系统的预测状态表示的新算法. 首先通过 suffix-history 算法, 发现系统的 landmark 的集合, 然后证明了任意 landmark 均可作为系统的初始状态. 从而, 可确定任意时间步下系统所处的经历, 进而估计任意经历下任意检验发生的概率, 发现和学习不可复位动态系统的预测状态表示.

本文假定所研究的系统是离散、观测值有限、且可用有限 POMDP 表示的随机系统.

2　PSRs 介绍

上面已经提到, PSR 模型采用检验核中所有检验发生概率所组成的向量作为其状态表示. 假定当前系统的观测值集合为 $O=\{o^1, o^2, \cdots, o^{|O|}\}$, 动作集合为 $A=\{a^1, a^2, \cdots, a^{|A|}\}$, 则长度为 n 的经历 $h=\{a^1o^1a^2o^2\cdots a^no^n\}$ 发生的概率是从初始时刻开始, 采取动作序列 $a^1a^2\cdots a^n$ 后, 观测值序列 $o^1o^2\cdots o^n$ 出现的概率, 即 $p(h)=prob(o_1=o^1, o_2=o^2, \cdots, o_n=o^n \mid a_1=a^1, a_2=a^2, \cdots, a_n=a^n)$. 同样, 长度为 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: $\boldsymbol{p}(t \mid h)=p(ht)/p(h)=prob(o_{n+1}=o^1, o_{n+2}=o^2, \cdots, o_{n+m}=o^m \mid h, a_{n+1}=a^1, a_{n+2}=a^2, \cdots, a_{n+m}=a^m)$. 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 后在该时刻出现的观测值. 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 $\boldsymbol{p}(Q \mid h)=[p(q_1 \mid h), p(q_2 \mid h), \cdots, p(q_k \mid h)]^{\mathrm{T}}$ 是所有经历的充分统计量, 即 $\boldsymbol{p}(Q \mid h)$ 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: $p(t \mid h)=f_t(\boldsymbol{p}(Q \mid h))$, 则认为检验集合 Q 构成一个PSRs. 其中 $\boldsymbol{p}(q_i \mid h)$ 是在经历 h 下检验 q_i 发生的概率. Q 被称为检验核, $\boldsymbol{p}(Q \mid h)$ 被称为预测向量, 以 $\boldsymbol{p}(Q \mid h)$ 作为 PSRs 的状态表示. 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.

在数学上, 可用系统动态矩阵 $\boldsymbol{Z}$ 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 $\boldsymbol{Z}$ 推导出来[2]. 如图 1 所示, $\boldsymbol{Z}$ 的行对应所有可能的经历(过去), 其中第一行对应空经历(null history) Φ 即初始经历, 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.

	t_1	…	t_j	…
$h_1=\phi$	$p(t_1 \mid h_1)$	…	$p(t_j \mid h_1)$	
h_2	.			
⋮	.			
h_i	$p(t_1 \mid h_i)$	…	$p(t_j \mid h_i)$	
⋮	⋮			

图1　系统动态矩阵 $\boldsymbol{Z}$

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 $\boldsymbol{Z}$ 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

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

长度为 k 的权向量 m_t, 使得相应于检验 t 的矩阵的列 $p(t|h)$, 可表示为 $p(t|h)=p(Q|h)^{\mathrm{T}}m_t$. 该式表明, 在得到经历 h 的预测向量 $p(Q|h)$ 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公式(1)进行计算或更新. 即对 $\forall q_i\in Q$[3]:

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

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[3]:

$$p(Q|hao)=\left(\frac{p(Q|h)^{\mathrm{T}}M_{ao}}{p(Q|h)^{\mathrm{T}}m_{ao}}\right)^{\mathrm{T}} \quad (2)$$

其中 M_{ao} 是一个 $k\times k$ 矩阵, 第 i 列对应的是 m_{aoq_i}. 从式(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 $p(Q|\phi)$ 后, 通过计算可得到任意经历的预测向量, 即可得知任意经历下系统所处的状态.

3 不可复位动态系统的预测状态表示的发现和学习

发现和学习受控动态系统的预测状态表示, 首先需要通过训练数据获取系统动态矩阵的任意元素, 即任意经历 h 下任意检验 t 发生的概率 $p(t|h)$. 对于可复位的动态系统, 由于可以得到一组从初始状态开始的训练数据, 经历/检验对 h/t 可能多次出现, 因此 $p(t|h)$ 可以通过蒙特卡罗方法估计得到[7]:

$$p(t|h)=\frac{success(t|h)}{exec(t|h)} \quad (3)$$

其中, $exec(t|h)$ 表示的是在经历 h 下, 检验 t 的动作序列发生的次数, $succeed(t|h)$ 表示的是在经历 h 下, 检验 t 发生的次数. 但对于不可复位的系统, 仅能得到一个连续的动作-观测值对序列, 对于任意经历, 任意检验在该经历下发生的次数为 1, 因此无法利用公式(3)估计 $p(t|h)$ 的值.

3.1 预备知识

文献[8]提出了记忆、landmark 的定义及性质.

记忆(memory): 动作、观测交替出现的序列, 结束于观测. 与经历不同的是, 长度为 n 的记忆表示该记忆中动作和观测的总数目为 n; 而长度为 n 的经历, 则是动作-观测对的数目为 n, 并且记忆不一定起始于动作或初始时刻.

Landmark: 构建系统动态矩阵 Z 的一个子矩阵, 其行对应的经历的集合为结束于某一个记忆的所有经历, 其列对应于 Z 的检验列. 如果该矩阵的秩为 1, 则该记忆为 landmark.

Landmark 的性质: Landmark 是经历的充分统计量, 可以作为系统的状态.

在可以精确获得系统中任意经历 h 下任意检验 t 发生的概率 $p(t|h)$ 的前提下, 并假定当前系统是可遍历的, 本文提出如下定理及推理.

定理 1 以不同的 PSR 状态作为系统的初始状态所得到的系统的 PSR 模型, 除了初始预测向量不同以外, 其余参数均相同.

证明: 假定 D、D' 分别为以不同的 PSR 状态作为系统的初始状态得到的系统动态矩阵, D 和 D' 对应的模型参数分别为 M_{ao}, m_{ao}, M'_{ao}, m'_{ao}. 由于 D 和 D' 的行对应系统中所有可能发生的 PSR 状态, 而对应同一个 PSR 状态的不同经历的预测向量相同, 故 D 和 D' 包含相同的行向量(但同一个行向量在 D 和 D' 的位置有可能不同). 在 D 和 D' 选择同一个检验核的前提下, 对于任意检验 t, $m_t=m'_t$. 故 $M_{ao}=M'_{ao}$, $m_{ao}=m'_{ao}$.

在定理 1 的证明中, 由于检验核的选择不影响系统的预测状态表示, 因此可假定不同的系统动态矩阵选择同一个检验核.

由 landmark 的性质可得, landmark 可作为系统的 PSR 状态, 而定理 1 表明了任意 PSR 状态均可作为系统的初始状态. 因此, 可得如下推理.

推理 1 任意 landmark 均可作为系统的初始状态.

由公式(3)可知, 对于一个连续的动作观测对序列, 如果可以确定该数据中每一时间步下系统所处的经历, 则 $p(t|h)$ 可根据蒙特卡罗方法估计得到. 为此, 本文首先提出任意时间步下经历的确定方法.

3.2 任意时间步下经历的确定方法

在给出了 landmark 的发现方法的基础上, 如表 1.

表 1 发现 landmark 的伪代码

初始设定:

d ←训练数据; m_t ←检验的最大长度; m_h ←经历的最大长度; m_i ←所有长度为 i 的记忆的集合; $H_m(m_h)$ ←结束于记忆 m, 长度为 m_h 的经历的集合; $T(m_t)$ ←长度不大于 m_t 的检验的集合; i ←1

算法流程:

for $m\in M_i$ /* 每一个属于 M_i 的记忆* /

根据数据 d 和 suffix-history 算法获取矩阵 $p(T(m_t)|H_m(m_h))$[5], 并计算 $p(T(m_t)|H_m(m_h))$ 的秩 $rank(p)$

If $rank(p)==1$

m 为 landmark

endfor

If 对于所有 $m\in M_i$, $rank(p)\neq 1$ 或者想发现更多的 landmark

i ←$i+1$ 得到 M_i, 如果有些记忆已经被判断为是 landmark, 则 M_i 不包含结束于这些记忆的记忆

从 for 开始, 继续以上过程

else

算法结束, 返回 landmark 的集合 L

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

任意时间步下系统所处的经历的确定方式为: 首先, 根据推理 1, 选择训练数据中出现的第一个 landmark 作为系统的初始状态, 并将其对应的经历定义为空经历, 然后从该 landmark 开始, 以后的每一时间步的经历的确定方式如表 2 所示.

表 2　确定系统所处经历的伪代码

```
初始设定:
d ←训练数据; L ←系统的 landmark 的集合; Lstart ←数据
d 中第一次出现的 landmark
Lm ←非 Lstart 的 landmark
算法流程:
for  在 d 中从 Lstart 开始, 此后的每一步
  if
      到该步为止的动作、观测序列结束于 Lstart
      系统当前的经历为: 空经历 φ
  else if
      到该步为止的动作、观测序列结束于某 landmark Lm
    if
      Lm 在 d 中首次出现
      系统当前的经历为: 以在该步之前出现的、离该步最近的
      Lstart 到该步之间的动作-观测序列的联合作为该步的经历,
      并将该经历作为 landmark Lm 对应的经历.
    else
      系统当前的经历为: Lm 对应的经历
  else
      系统的当前经历为: 以在该步之前出现的、离该步最近的
      landmark 的经历和该 landmark 到该步之间的动作-观测序列的
      联合作为该步的经历, 例如, 如果在该步之前出现的、离该步
      最近的 landmark 的经历为 a1 o1 a2 o2, 该 landmark 到该步之间
      的动作-观测序列为 a3 o3 a4 o4, 则当前经历为
      a1 o1 a2 o2 a3 o3 a4 o4
endfor
```

从表 2 可以看出, 如果系统中存在 landmark 的集合, 则根据本文算法可以准确确定每一时间步系统所处的经历, 避免了 suffix-history 算法存在的很多情况下不能准确确定系统所处经历的问题.

3.3　发现和学习系统的预测状态表示

在任意时间步对应的经历可以确定后, 首先给出获取系统动态矩阵 $\boldsymbol{Z}$ 的任意子矩阵的方法, 如表 3 所示, 在此基础上发现和学习不可复位动态系统的预测状态表示的具体步骤为(由于通过估计得到的 $\boldsymbol{Z}$ 的子矩阵的元素存在噪声, 本文按照文献[7], 采用计算元素有噪声的矩阵的秩的方法来计算矩阵的秩):

步骤 1　根据表 3 即 $\boldsymbol{p}(T|H)$ 的获取方法, 获取系统动态矩阵 $\boldsymbol{Z}$ 的子矩阵 $\boldsymbol{Z}_1$, 其行对应的经历的集合是 d 中包含的所有经历, 其列对应的检验的集合为 $T_1=\{\forall a, o: ao\}$. 获取矩阵 $\boldsymbol{Z}_1$ 的秩 v_1, 检验核 $Q_{T_1}(\boldsymbol{Z}_1$ 中任意 v_1 个线性无关的检验的集合).

步骤 2　根据表 3, 获取系统动态矩阵 $\boldsymbol{Z}$ 的子矩阵 $\boldsymbol{Z}_i(i\geqslant 2)$, 其行对应的经历的集合是 d 中包含的所有经历, 其列对应的检验的集合为 $T_i=\{\forall a, o, t\in Q_{T_{i-1}}: ao, t, aot\}$. 获取矩阵 $\boldsymbol{Z}_i$ 的秩 v_i, 检验核 $Q_{T_i}(\boldsymbol{Z}_i$ 中任意 v_i 个线性无关的检验的集合). 如果 v_i 和 v_{i-1} 不相同, $i\leftarrow i+1$, 继续执行步骤 2; 否则, 即 $v_i==v_{i-1}$, 算法停止, 并以 Q_{T_i} 做为发现的系统的检验核 Q_T. 令 $k=v_i$, 同时获取系统的经历核 $Q_H(\boldsymbol{Z}_i$ 中任意 v_i 个线性无关的经历的集合).

然后, 根据公式(4)计算得到系统的 PSR 模型的各个参数[7]:

$$\boldsymbol{m}_{ao}=\boldsymbol{p}^{-1}(Q_T|\boldsymbol{Q}_H)\boldsymbol{p}(ao|Q_H)$$
$$\boldsymbol{m}_{aoq_i}=\boldsymbol{p}^{-1}(Q_T|\boldsymbol{Q}_H)\boldsymbol{p}(aoq_i|Q_H)\qquad(4)$$

表 3　获取矩阵 $\boldsymbol{p}(T|H)$ 的伪代码

```
初始设定:
d ←训练数据; H ←给定经历的集合; T ←给定检验的集合;
success(t|h) ←经历 h 下检验 t 发生的次数; exec(t|h) ←经历 h 下检
验 t 对应的动作序列发生的次数, 显然, 不同的检验对应的动作序列
可能相同; m_t ←检验集合 T 中检验的最大长度
算法流程:
for   d 的每一步 m
      根据本文提出的经历的确定方式即表 2, 确定当前步 m 的经
      历 h 对于每一个从当前步 m 开始, 并且长度不大于 m_t 的检
      验 t, 将其对应的动作序列在经历 h 下的发生次数加 1, 即 exec
      (t|h) ←exec(t|h)+1, 同时如果 t∈T, 令 success(t|h) ←suc-
      cess(t|h)+1
endfor
      矩阵 p(T|H) 中任意元素 p(t|h) 的值为:
      p(t|h) ←success(t|h)/exec(t|h)
```

4　实验

为了验证所提算法的性能, 将本文算法应用于一组标准的 POMDP 系统, 这些系统也是研究 POMDP 以及 PSRs 常用的测试系统, 关于这些系统的详细描述见文献[9]. 下面描述算法的实验内容与结果, 并且将本文算法和 suffix-history 算法作了比较分析.

针对每个系统, 智能体采取随机策略得到一训练数据后, 通过本文所提算法得到该系统的预测状态表示, 即系统的 PSR 模型. 其中, 为了简单起见, 用于获取系统的预测状态表示的训练数据是采用随机策略获取的, 当然也可用非随机策略获取[10]. 然后采用随机策略产生一测试数据, 并采用一种标准化的指标衡量得到的预测状态表示的准确性, 即在采取某一动作后, 首先根据得到的 PSR 模型计算所有观测值出现的概率, 然后根据其 POMDP 模型计算实际上所有观测值出现的概率, 通过两者之间的平方差判断 PSR 模型的准确性, 具体的说是通过公式(5)来判断[11]:

$$\frac{1}{L}\sum_{t=1}^{L}\frac{1}{|O|}\sum_{o\in O}(p(o|h_t, a_t)-\hat{p}(o|h_t, a_t))^2\qquad(5)$$

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

其中, L 表示的是产生的测试数据的长度, 在本文中 $L=10,000$. $p(o|h_t, a_t)$ 是根据系统的POMDP 模型计算得到的, 经历 h_t 下采取动作 a_t 产生观测值 o 的真实概率, $\hat{p}(o|h_t, a_t)$ 是通过 PSR 模型计算得到在经历 h_t 下采取动作 a_t 产生观测值 o 的概率, 即 $\hat{p}(o|h_t, a_t) = p(Q|h_t)\, m_{a_t o}$. 显然, 得到的平方差值越小表明 PSR 模型的预测精度越高, 算法的性能越优越.

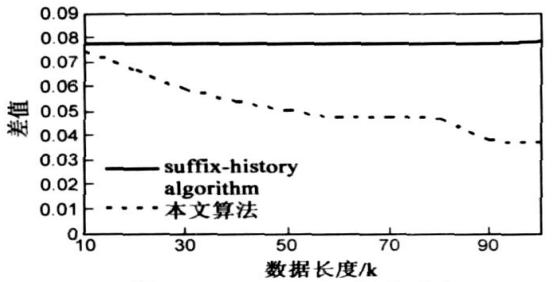

图2 Cheese Maze(奶酪迷宫)

对每个环境, 仅在长度为 1 的记忆中寻找 landmark. 同时, 为了增强算法的可比性, 两种算法采用相同的参数, 其中经历的最大长度为 3, 检验的最大长度为 2. 对于两种算法, 各做了 10 次试验. 对相同长度的训练数据, 以 10 次实验得到的平均值作为该长度对应的差值. 试验结果如图 2、3、4 所示, 其中横坐标表示的是用于获得 PSR 模型的训练数据的长度, 纵坐标表示的是采用公式(5) 计算得到的差值.

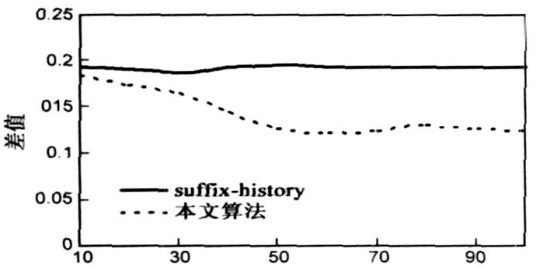

图3 Shuttle(穿梭机)

从图 2、3、4 可以看出, 针对三个不同的系统, 通过本文算法得到的差值均明显小于 suffix-history 算法得到的差值, 即本文算法获取的系统的 PSR 模型的预测精度明显优于 suffix-history 算法. 根据以上结果, 说明本文所提算法是有效的, 并且由于本文仅在长度为 1 的记忆中寻找 landmark, 如果增加寻找的长度, 发现系统中更多的 landmark 有可能会使得本文所提算法的效果更好, 有待进一步的实验验证.

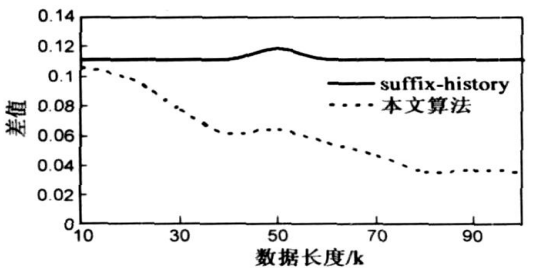

图4 4×3 Maze(4×3迷宫)

5 结论与展望

在对一个离散时间、有限观测值的受控动态系统建模时, 相比于其它两种常用的 POMDPs 和基于经历的方法, PSRs 具有很多优点, 因此得到了越来越多的重视. 不可复位动态系统的预测状态表示的发现和学习是针对 PSRs 研究的重要内容, 在本文中, 首先提出了发现系统的 landmark 的方法, 然后证明了系统的任意 landmark 均可作为系统的初始状态, 从而可利用 landmark 的性质确定系统在每一时间步所处的经历, 进而提出发现和学习不可复位动态系统的预测状态表示的一种新算法. 算法被应用于一些标准的 POMDP 系统, 经过仿真, 结果表明, 本文算法得到的系统的 PSR 模型在预测精度上明显优于 suffix-history 算法.

对 PSRs 的研究可以说仍然处在起步阶段, 很多问题仍然没有得到解决, 例如, 如何减少训练的数据量; 如何将 PSRs 应用于状态空间、动作空间连续的系统以及比较大的、复杂的系统; 如何将学习到的一个简单系统的 PSR 模型转移到一个更复杂的相似的系统中等. 针对这些问题的研究将是以后研究的重点.

参考文献:

[1] L Kaelbling, M Littman, A Cassandra. Planning and acting in partially observable stochastic domains[J]. Artificial Intelligence, 1998, 101(1-2): 99-134.

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

[3] M Littleman, R Sutton, S Singh. Predictive representation of state[A]. In Advances in Neural Information Processing Systems 14[C]. Vancouver, British Columbia, Canada: MIT Press, 2002. 1555-1561.

[4] M Rosencrantz, G Gordon, S Thrun. Learning low dimensional predictive representations[A]. In Proceedings of the Twenty-First International Conference on Machine Learning[C]. Banff, Alberta, Canada: ACM, 2004. 695-702.

[5] B Wolfe, M James, S Singh. Learning predictive state representations in dynamical systems without reset[A]. In Proceedings of the Twenty-Second International Conference on Machine Learning[C]. Bonn, Germany: ACM, 2005. 980-987.

[6] D Wingate, S Singh. On discovery and learning of models with predictive state representations of state for agents with continu

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

ous actions and observations[A]. In Proceeding of the 2007 International Conference on Autonomous Agents and Multiagent Systems[C]. Honolulu, USA: ACM, 2007. 187- 194.

[7] M James, S Singh. Learning and discovery of predictive state representations in dynamical systems with reset[A]. In Proceedings of the Twenty-First International Conference on Machine Learning[C]. Banff, Alberta, Canada: ACM, 2004. 417-424.

[8] M James, B Wolfe, S Singh. Combining memory and landmarks with predictive state representations[A]. In Proceedings of the International Joint Conference on Artificial Intelligence[C]. Edinburgh, Scotland: Professional Book Center, 2005. 734- 739.

[9] A Cassandra. Tony' s POMDP file repository page [OL]. http://www. cs. brown. edu/research/ai/pomdp/examples /index . html, 2008- 06- 02.

[10] M Bowling, P McCracken, M James, et al. Learning predictive state representations using non-blind policies[A]. In Proceedings of the Twenty-Third International Conference on Machine Learning[C]. Pittsburgh, Pennsylvania, USA: ACM, 2006. 129-136.

[11] S Singh, M Littman, N Jong, et al. Learning predictive state representations[A]. In Twentieth International Conference on Machine Learning[C]. Washington, DC, USA: AAAI Press, 2003. 712- 719.

作者简介:

刘云龙　男, 1977 年 11 月出生于山东安丘, 西安交通大学博士研究生. 研究方向为强化学习、预测状态表示、不确定环境下的规划等.
E-mail: ylliusv@ 163. com

李人厚　男, 1935 年 5 月出生于浙江宁波, 教授, 博士生导师. 研究方向为智能控制理论与方法、预测状态表示、CSCW 理论与应用等.
E-mail: rhli@ mail. xjtu. edu. cn

(上接第 117 页)

[8] K Rahbar, J P Reilly. A frequency domain method for blind source separation of convolutive audio mixtures[J]. IEEE Transactions on Speech and Audio Signal Processing, 2005, 13(5): 832-844.

[9] Liang J, Ding Z. Blind MIMO system identification based on cummulant subspace decomposition[J]. IEEE Transactions on Signal Processing, 2003, 51(6): 1457- 1468.

[10] Marc Castella, Saloua Rhioui, Eric Moreau, et al. Quadratic higher order criteria for iterative blind separation of a MIMO convolutive mixture of sources[J]. IEEE Transactions on Signal Processing, 2007, 55(1): 218- 232.

[11] Feng D Z, Zhang X D, Bao Z. An efficient multistage decomposition approach for independent components[J]. Signal Processing, 2003, 83(1): 181-197.

[12] LaSalle J P. The Stability of Dynamical Systems (first edition)[M]. Philadelphia PA: IAM Press, 1976. 5- 11.

[13] A Belouchrani, K Abed-Meraim, J F Cardoso, et al. A blind source separation technique using second-order statistics[J]. IEEE Transactions on Signal Processing, 1997, 45(2): 434-444.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net