一种基于模型的可分解贝叶斯在线强化学习

仵 博[1],郑红燕[1],冯延蓬[1],陈 鑫[2,3]
(1.深圳职业技术学院教育技术与信息中心,广东深圳 518055;2.中南大学信息科学与工程学院,湖南长沙 410083;
3.先进控制与智能自动化湖南省工程实验室,湖南长沙 410083)

摘 要: 针对贝叶斯强化学习中参数个数巨大,收敛速度慢,无法实现在线学习的问题,提出一种基于模型的可分解贝叶斯强化学习方法.首先,将学习参数进行可分解表示,降低学习参数的个数;然后,根据先验知识和观察数据采用贝叶斯方法来学习,最优化探索和利用二者之间的平衡关系;最后,采用基于点的贝叶斯强化学习方法实现学习过程的快速收敛,从而达到在线学习的目的.仿真结果表明该算法能够满足实时系统性能的要求.

关键词: 马尔可夫决策过程; 贝叶斯强化学习; 动态贝叶斯网路
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112 (2014)07-1429-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2014.07.029

Model-Based Factored Bayesian Online Reinforcement Learning

WU Bo[1], ZHENG Hong-yan[1], FENG Yan-peng[1], CHEN Xin[2,3]
(1. *Education Technology and Information Center, Shenzhen Polytechnic, Shenzhen, Guangdong* 518055, *China*;
2. *School of Information Science and Engineering, Central South University, Changsha, Hunan* 410083, *China*;
3. *Hunan Engineering Laboratory for Advanced Control and Intelligent Automation, Changsha, Hunan* 410083, *China*)

Abstract: Due to the enormous number of parameters and slow convergence which are the major obstacles for online learning in model-based Bayesian reinforcement learning, the paper presents a model-based factored Bayesian reinforcement learning approach. Firstly, factored representations are made to represent the dynamics with fewer parameters. Then, according to prior knowledge and observable data, this paper exploits model-based reinforcement learning to provide an elegant solution to the optimal exploration-exploitation tradeoff. Finally, a pointed-based Bayesian reinforcement learning approach is proposed to speed up the convergence to achieve online learning. The experimental results show that the proposed approach can approximate the underlying Bayesian reinforcement learning task well with guaranteed real-time performance.

Key words: Markov decision processes; Bayesian reinforcement learning; dynamic Bayesian networks

1 引言

在实际的动态系统中,状态转移函数通常是未知的,且为动态变化的.如果使用稳态模型描述动态系统,则会造成对动态系统建模的失真,在理论上也无法保证获得真实近似最优值函数和最优策略.有鉴于此,需要智能体在与动态不确定环境交互中进行学习.强化学习是一种有效的最优控制学习方法,可在模型复杂或者不确定等条件下,实现系统基于数据驱动进行多阶段优化学习控制[1,2].

经典增强学习算法按照是否基于模型分类,可分为基于模型(Model-Based)和模型自由(Model-Free)两类.基于模型的有 TD 学习、Q 学习和 SARSA[3]等算法.模型自由的有 DYNA-Q 和优先扫除等算法.以上经典增强学习算法在理论上证明了算法的收敛性,在实际的应用领域,学习的参数个数很多,是一个典型的 NP 难问题,难以最优化探索和利用两者的平衡[4,5].

贝叶斯强化学习(Bayesian Reinforcement Learning, BRL)利用模型先验知识对未知模型参数进行建模,然后根据观察数据对未知模型参数的后验分布进行更新,最后根据后验分布进行规划,以获得最大化期望报酬值.BRL 本质上是将学习问题转化为规划问题.由于 BRL 能够利用状态的先验分布对未知参数和未知模型进行建模,为最优化探索和利用的平衡提供一种完美的解决方案,得到国内外学者的广泛关注,成为当前强化学习领域研究的热点[6,7].但 BRL 存在两个难题:一是学习参数个数巨大,并呈指数规模增长[8];二是在全部后验信念状态空间上求解规划问题将遭遇"维数灾"[9].以上两个难题使得现有的 BRL 算法只能够求解小规模问题,无法实现大规模问题的在线学习.

收稿日期:2013-08-30;修回日期:2014-03-04;责任编辑:孙瑶
基金项目:国家自然科学基金(No.61074058,No.60874042);深圳市自然科学基金(No.JCYJ20120617134831736)

本文针对以上问题,提出一种基于模型的可分解贝叶斯强化学习方法.采用可分解表示方法,降低学习参数的规模.针对 DBNs(Dynamic Bayesian Networks, DBNs)结构(变量之间的独立关系)未知情况,采用贝叶斯方法对未知结构和参数同时进行学习.最后采用基于点的贝叶斯强化学习(Pointed-Based BRL, PB-BRL)方法在后验信念空间进行动作选择,实现在线规划和学习.实验和仿真结果表明,本文算法能够有效降低参数个数,实现对动态系统的在线学习.

2 贝叶斯强化学习建模

马尔可夫决策过程(Markov Decision Processes, MDPs)可以用四元组 $<S,A,T,R>$ 来描述.状态集合 $S=\{s_1,s_2,\cdots,s_n\}$,包含智能体所有可能的处在的状态;动作集合 $A=\{a_1,a_2,\cdots,a_m\}$,包含智能体所有可能的采取的动作;状态转移函数 $T(s,a,s')=P(s'\mid s,a)$,当智能体在状态 s 下采用动作 a,转移到状态 s' 的概率;报酬函数 $R(s,a,s')$,在状态 s 下采用动作 a 转移到状态 s' 获得的报酬值.

在强化学习中,状态转移函数 $T(s,a,s')$ 为未知学习参数 $\theta^{sas'}$.根据文献[10],将基于模型的贝叶斯强化学习定义为部分可观察马尔可夫决策过程(Partially Observable Markov Decision Processes, POMDPs),用六元组 $<S_P,A_P,Z_P,T_P,O_P,R_P>$ 描述.其中,S_P 是离散的状态 S 与连续的未知参数 $\theta^{sas'}$ 的叉积;动作集合 A_P 与 MDPs 的动作集合 A 相同;$Z_P=S$.状态转移函数 $T_P(s,\theta,a,s',\theta')=P(s',\theta'\mid s,\theta,a)$ 可被分解为两个条件分布的乘积:

$$\begin{aligned}T_P(s,\theta,a,s',\theta')&=P(s',\theta'\mid s,\theta,a)\\&=P(s'\mid s,\theta,a,\theta')P(\theta'\mid s,\theta,a)\\&=\theta^{sas'}\delta^{\theta\theta'}\end{aligned}\tag{1}$$

其中,$\delta^{\theta\theta'}$ 为克罗内克函数,满足:

$$\delta^{\theta\theta'}=\begin{cases}1, & \theta'=\theta\\0, & \text{otherwise}\end{cases}\tag{2}$$

观察函数 $O_P(s',\theta',a,z)=P(z\mid s',\theta',a)$ 表示为当智能体执行动作 a,状态和参数转移到 s' 和 θ' 时,观察为 z 的概率.由于报酬函数不依赖于 θ 和 θ',报酬函数 $R_P(s,\theta,a,s',\theta')=R(s,a,s')$ 与 MDPs 的报酬函数相同.

根据贝叶斯强化学习的定义,将 MDPs 问题转化为 POMDPs 问题.在 POMDPs 中,由于状态是未知的,引入状态 S 的概率分布 $b(s)$,称之为信念.通过信念概念的引入,θ 可以采用信念监控方法来学习[11].利用贝叶斯更新法则,信念 $b(\theta)$ 的更新如下:

$$b^{sas'}(\theta)=\eta b(\theta)P(s'\mid\theta,s,a)=\eta b(\theta)\theta^{sas'}\tag{3}$$

其中,η 为归一化因子.

只有当信念的先验和后验分布处在同一分布族时,信念监控方法才有效.贝叶斯强化学习采用狄利克雷(Dirichlet)分布来表示信念先验和后验分布.假设信念先验为 $b(\theta)=\prod_{s,a}D(\theta^{sa};\boldsymbol{n}^{sa})$,$nsa$ 为超参数 $n^{sas'}$ 的向量,则其后验分布为:

$$\begin{aligned}b^{sas'}(\theta)&=\eta\theta^{sas'}\prod\nolimits_{s,a}D(\theta^{sa};n^{sas'})\\&=\prod\nolimits_{s,a}D(\theta^{sa};\boldsymbol{n}^{sa}+\delta_{\tilde{s},\tilde{a},\tilde{s}'}(s,a,s'))\end{aligned}\tag{4}$$

其中,δ 仍为克罗内克函数,当 $s=\tilde{s}$、$a=\tilde{a}$ 和 $s'=\tilde{s}'$ 时为 1,其他为 0.

强化学习的目标是根据当前模型的后验和状态,在状态转移函数未知情况下寻找最优策略来最优化平衡探索和利用,以获得最大的长期报酬值.在 POMDPs 中,策略 π 是信念 b 到 a 的映射,即 $\pi(b)\rightarrow a$.最优策略 π^* 为最优值函数 V^* 对应的策略.

3 基于模型的可分解贝叶斯强化学习

为了保证在模型不确定的情况下仍能获得一个好的模型,需要在学习过程中收集大量的数据,造成学习参数呈指数级爆炸性增长,导致 BRL 无法实现快速收敛.可分解表示法是解决学习参数"维数灾"的有效方法[12].在可分解表示方法中,如果 DBNs 中变量之间的独立关系已知,学习参数规模可以很容易实现压缩.而在实际的应用领域中,DBNs 的结构是未知.因此,需要对 DBNs 的结构和参数同时进行学习.

3.1 可分解学习表示

在大多数真实世界模型中,通过分析状态变量内部结构可以发现,状态变量可以用一组随机变量集合来表示,这种内部特性称为可分解特性.可分解状态变量用一个随机有限变量集合 $X=\{X_1,\cdots,X_n\}$ 来表示,每一个 X_i 表示状态变量的一个特征,X_i 表示集合 X 中每个变量的取值集合,一个状态可以表示成为 $s=\{X_1=x_1,\cdots,X_n=x_n\}$,其中 $x_i\in X_i$,也可以简化表示为 $s=\{x_i\}_{i=1}^n$,状态变量的空间为 $|S|=\prod_{i=1}^n|X_i|$.对状态变量进行可分解后,状态转移函数、观察函数以及报酬函数都可以使用 DBNs 来压缩表示[12].

定义 $G(a)$ 为一个两层的有向无环图,其中,$a\in A$,结点为 $X=\{X_1,\cdots,X_n,X_1',\cdots,X_n'\}$.定义 $\theta_{G(a)}$ 为条件概率表,则状态转移函数 T 可由 $G(a)$ 和 $\theta_{G(a)}$ 来表示.定义 X_i 为当前状态下的第 i 个特征变量,X_i' 为下一时刻状态的第 i 个特征变量.$X_{\prec i}^{G(a)}$ 为特征变量 $X_i'=x_i$ 时父亲结点的取值.状态转移函数计算如下:

$$T(s,a,s')=T(X,a,X')=\prod_{i=1}^{n}P(X_i'\mid X_{\prec i}^{G(a)},a)\tag{5}$$

其中,$T(s,a,s')$为未分解表示时的状态转移函数,$T(X,a,X')$为可分解表示时的状态转移函数.

在 FMDPs(Factored MDPs,FMDPs)中,由于状态转移函数、观察函数和报酬函数都可以采用 DBNs 的条件概率表来表示,因此,可以通过重复计算信念后验来同时学习以上未知模型.

3.2　信念后验更新

可分解贝叶斯强化学习是通过智能体与环境的交互,获取观察数据,学习未知参数 $\theta_{G(a)}$和未知结构$G(a)$,从而建立状态转移模型和报酬模型.在确定性环境下,给定初始信念 $b(s)$,其信念后验 $b_{a,z'}(s')$计算如下:

$$b_{a,z'}(s') = \eta\delta([s']_{Z'} = z')\sum_{s} b(s)P(s' \mid s,a) \tag{6}$$

其中,η 为归一化常量;$[s']_{Z'}$为观察变量集合 Z'对应的状态变量值的子集;δ 为克罗内克函数,当$[s']_{Z'} = z'$为真时,返回值为 1;为假时,返回值为 0.由于模型和结构未知,根据上节的知识,可知信念状态的更新过程如下:

$$b(X',\theta_{G(a)}) = \eta\delta\sum_{X} P(X' \mid X,a,\theta_{G(a)})b(X,\theta_{G(a)}) \tag{7}$$

其中,X 和 X'为可分解表示的变量特征;a 为动作;Z 为观察数据集合,z 为在 Z 上的子集;$\theta_{G(a)}$为未知参数;δ 为克罗内克函数.

但是,由于信念状态更新需要历史信息,需要遍历所有的历史观察和动作,造成式(7)无法收敛.根据文献[9]的结论可知,信念状态更新在 Dirichlet 混合乘积上是闭环的.因此,可以采用 Dirichlet 混合乘积来表示信念状态.信念状态先验概率的 Dirichlet 混合乘积表示形式为:

$$b(X,\theta_{G(a)}) = \sum_{i} c_{i,X}\prod D_{i,X}(\theta_{G(a)}^{X_{<i}^{C}}) \tag{8}$$

其中,$c_{i,X}$为 Dirichlet 系数;D 为 Dirichlet 分布函数;$\theta_{G(a)}^{X_{<i}^{C}} = P(X' \mid \text{parents}(X'))$.则,信念状态更新后的后验信念为:

$$b_{a,z'}(X',\theta_{G(a)}) = \sum_{j} c_{j,X'}\prod D_{j,X'}(\theta_{G(a)}^{X_{<j}^{C}}) \tag{9}$$

3.3　值函数参数化

根据上文知识可知,在 FMDPs 领域中的贝叶斯强化学习可由带模型变量 $\theta_{G(a)}$的 DBNs 来建模.如果把未知模型变量 $\theta_{G(a)}$作为 FMDPs 的隐变量,可将带有未知参数的 FMDPs 转化为 FPOMDPs(Factored POMDPs,FPOMDPs).根据以上结论,就可以采用现有 FPOMDPs 规划算法来求解 BRL 问题.

Poupart 等人提出一种基于点的值迭代强化学习的 BEETLE 算法[9],该算法只适用于 MDPs 领域.在此基础上,Porta 等人[13]提出一种改进型 BEETLE 来处理连续状态的 POMDPs 贝叶斯强化学习.BEETLE 算法和其改进型算法都充分利用这样一个事实:最优值函数都是 α-函数集合上界面的参数化形式,α-函数为多元多项式.但是,它们是基于 MDPs 或 POMDPs 的,无法泛化到 FPOMDPs 领域[14].本节借鉴 BEETLE 思想,将值函数参数化拓展到 FPOMDPs 领域.

离散 POMDPs 的最优值函数具有分段线性凸特性,即最优值函数可以用线性分段集合 $\boldsymbol{\Gamma}$(称之为 α-向量)的上界面来表示,公式化描述为:

$$V^{*}(b) = \max\ \alpha(b) \tag{10}$$

每个 α 是每个特征变量概率值的线性组合,即 $\alpha(b) = \sum_{X} c_X b(s)$.对于离散状态空间,状态个数有界,$\alpha$ 可以表示成每个特征变量的 Dirichlet 系数向量,即$\alpha(X) = c_X$.对于连续状态 POMDPs,最优值函数是线性函数(α-函数)集合 $\boldsymbol{\Gamma}$ 的上包络,公式化描述为:

$$\alpha(b) = \int_{X} c_X b(X)\mathrm{d}X \tag{11}$$

在可分解强化学习中,假设 k 时刻的最优值函数为 $V^k(b)$,α-函数集合为Γ^k,则:

$$V^k(b) = \max_{\alpha\in\Gamma^k}\alpha(b) \tag{12}$$

根据 Bellman 更新方程,$k+1$ 时刻的最优值函数为 $V^{k+1}(b)$,α-函数集合为 $\boldsymbol{\Gamma}^{k+1}$.由于引入了 α-函数,Bellman 更新方程可改写为:

$$V^{k+1}(b) = \max_{a\in A}\sum_{X} b(X)R(X,a,\theta_{G(a)}) + \gamma\sum_{z'} P(z' \mid b,a,\theta_{G(a)})\max_{\alpha\in\Gamma^k}\alpha(b_{a,z'}) \tag{13}$$

根据文献[7]的证明可知,α-函数是 Dirichlet 乘积的线性组合.在每步的 Bellman 备份中,线性组合中 Dirichlet 乘积个数等于状态空间大小.因此,α-函数的线性组合大小会随着决策时间呈指数规模增长.

3.4　基于点的贝叶斯强化学习算法

针对 α-函数的线性组合呈指数规模增长的问题,本文采用一种采用基于点的贝叶斯强化学习(PB-BRL)来实现在线的规划和学习.

PB-BRL 算法:输入为当前信念状态 b、与或树 T 的深度 d 和动态贝叶斯网络结构 $P(X_i \mid \text{parents}(X_i))$.其中,$X$ 为原始特征集合;A 为动作集合;S 为所有直接影响报酬的变量 X_i 的集合,$X_i \in X$;D 为最大扩展深度;a_{best}为最佳动作;$R_{\max}(b)$为最大报酬值;R_c 为当前报酬值;$U_T(b)$为报酬值上界;b_c 为当前信念状态点.输出为 $R_{\max}(b)$.

Step 1　采用动态贝叶斯网络对学习模型进行可分解表示,对于所有的 $X_i \in S$、$X_j \in X$ 和 $a \in A$,如果 $X_j \in \text{parents}(X_i') \wedge X_j \notin S$,则 $S \leftarrow S \cup S'$;

Step 2 当与或树 T 的深度 $d=0$ 时,计算边缘结点的最大报酬值,$\max_{a\in A}\sum_{s\in S}b(s)R(b,a)$,并将计算结果赋值给 R_c,$R_c\leftarrow R_B(b,a)$;

Step 3 对于每一个 $a\in A$,循环执行 Step 4、Step 5 和 Step 6;

Step 4 将 $R(s,a)$赋值给一个临时变量 R_{temp},即 $R(s,a)\rightarrow R_{\text{temp}}$;

Step 5 对于变量 k 从 1 到 N,执行 Monte Carlo 采样操作:首先从 $P(s'\mid s,b,a)$中采样 s',然后利用式(9)实现信念后验的更新,获得 b',实现贝叶斯学习,最后更新 $R_{\text{temp}}=R_{\text{temp}}+\frac{\gamma}{N}V(s',b',d-1,N)$;

Step 6 如果 $R_{\text{temp}}>R_{\max}(b)$,则将临时变量 R_{temp} 赋值给 $R_{\max}(b)$,将当前动作 a 赋值给 a_{best};

Step7 如果$(d=D)\cup\|V^*(b)-R_{\max}(b)\|<\varepsilon$ 成立,终止条件满足,则获得最佳动作 $a_{\text{best}}\leftarrow a$.

PB-BRL 算法是一种在线算法,Step 1 采用可分解表示,假设所有变量的父亲结点数都是 k,并且 $k\ll n$,那么需要遍历的参数规模为 $O(n2^k)$.如果利用变量的独立关系,使用决策树来表示,决策树上的叶子节点数不会超过 $f(k)$个,$f(k)$代表变量父亲节点数的多项式函数,那么需要遍历的参数规模为 $O(nf(k))$. Step 2 至 Step 7 是基于点的在线值迭代算法[15],由于变量信念与或树的最大深度为 D,因此其时间复杂度和空间复杂度都为 $O((|A||Z|)^D|\boldsymbol{\Gamma}||B|)$.由于采用基于点的值迭代方法,$|\boldsymbol{\Gamma}|$和$|B|$不会随着更新时间而成指数增长,一般情况下,它是一个固定值.因此,PB-BRL 算法时间复杂度为 $O(nf(k)+(|A||Z|)^D|\boldsymbol{\Gamma}||B|)$.

4 实验结果

针对 Chain 问题,与文献[9]的 BEETLE 算法和文献[16]的 MC-BRL 算法进行比较,这两个算法是最近几年提出的,能够代表当前贝叶斯强化学习算法的水平.针对小车爬山问题,与文献[18]的 EFSL 算法进行比较.

4.1 Chain 问题仿真实验

Chain 问题中,有两个动作$\{a,b\}$,五个状态$\{1,2,3,4,5\}$,动作的转移概率为 $P=0.2$.一旦到达状态 5,得到的报酬值为 10. Chain 有 Chain_Tied、Chain_Semi 和 Chain_Full 等三个版本. Chain_Full 是指状态转移函数 $T(s,a,s')$和状态转移结构 G 都未知. Chain_Semi 指的是状态转移结构已知,状态转移函数未知,动作之间存在依赖关系. Chain_Tied 是动态系统已知,动作的转移概率未知,动作和状态是独立的.因此,Chain 问题具有多样的不确定性,是评价强化学习算法的理想平台.

本文对 Chain 问题的三个版本分别进行 500 次实验,每次实验进行 1000 步(迭代次数),对于报酬值取实验结果的平均值和标准差,报酬值越大,算法越好.表 1 为不同算法报酬值的比较,其中,$n.v.$表示不可获得;Optimal 表示在理想状态下的最优值;BEETLE 为基于点的值迭代算法,该算法采用 DDNs(Dynamic Decision Networks,DDNs)来分解状态;MC-BRL 为基于 Monte Carlo 的贝叶斯强化学习算法;Q-Learning 是一种 ε-贪婪策略[17],ε 的取值范围为 0 到 0.5;PB-BRL 为本文算法.本文实验数据为 $K=1000$ 的采样结果.

表 1 不同算法报酬值的比较

问题	BEETLE	MC-BRL	Q-Learning	PB-BRL
Chain_Tied	3650 ± 41	3618 ± 29	1616 ± 24	3659 ± 20
Chain_Semi	3648 ± 41	$n.v.$	1616 ± 24	3661 ± 21
Chain_Full	1754 ± 42	1646 ± 32	1616 ± 24	2565 ± 23

从表 1 中的实验数据可知,在不确定因素较少的 Chain_Tied 和 Chain_Semi 问题中,PB-BRL 与 BEETLE 和 MC-BRL 的平均报酬值相差不大,但 PB-BRL 算法更接近真实最优值.在较大规模的 Chain_Full 问题中,状态转移函数和状态转移结构都未知,不确定因素较多,PB-BRL 的平均报酬值为 2565,明显高于 BEETLE 和 MC-BRL 算法,因此 BEETLE 具有更好的性能.因为,本文采用 Monte Carlo 采样方法可以有效地降低问题的求解规模,从而能够更好地平衡探索和利用.由于 Q-Learning 是这一种模型自由的强化学习方法,它与状态转移函数等模型独立,因此它的报酬值在 Chain 的三种版本中保持不变. BEETLE、MC-BRL 和 PB-BRL 是三种不同类型的贝叶斯强化学习方法,从表 1 可知,由于贝叶斯强化学习方法充分利用了先验知识,可以有效地增强学习效果,提高报酬值.

图 1 为 BEETLE、MC-BRL、Q-Learning 和 PB-BRL 四种算法的积累报酬值随着迭代次数的增加而变化的情况,本文实验为前 1000 步的迭代结果.从图 1 可知,本文 PB-BRL 算法的积累报酬值最大,而 Q-Learning 算法的积累报酬值最小,BEETLE 算法和 MC-BRL 算法的结果与 PB-BRL 算法结果接近.从积累报酬值的对比实验可知,PB-BRL 算法性能最好,贝叶斯强化学习较 Q 学习具有更好的性能.在前 1000 步的迭代中,算法的学习率为常数.

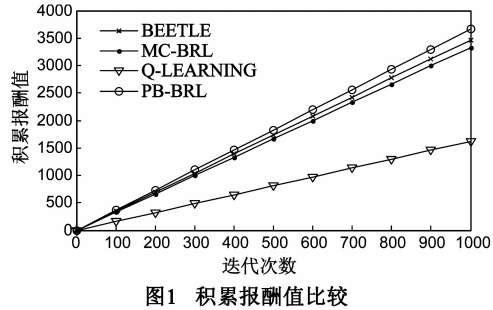

图1 积累报酬值比较

表 2　算法计算时间的比较(单位为 ms)

问题	BEETLE		MC-BRL		PB-BRL	
	离线	在线	离线	在线	离线	在线
Chain_Tied	400	1500	1.8e+6	32	400	18
Chain_Semi	1300	1300	*n.v.*	*n.v.*	1300	22
Chain_Full	14800	18000	*n.v.*	*n.v.*	14800	37

表 2 为不同贝叶斯强化学习算法计算时间的比较,其中,*n.v.*表示不可获得.从表中数据可知,PB-BRL 和 MC-BRL 在线计算时间上耗时更少,具有更高的实时性.但是,从表 1 的报酬值上看,MC-BRL 针对大规模问题求解时误差较大.PB-BRL 的离线计算方法与 BEETLE 相同,存在耗时问题.离线预计算不会影响算法的在线实时性;同时,离线训练可以获得更优的先验知识,从而可以获得尽可能大的积累报酬值,较好地解决强化学习中探索和利用这一难题.

4.2　小车爬山问题仿真实验

小车爬山学习控制在有关强化学习的文献中,通常被作为一个典型的连续状态空间强化学习问题来验证算法的学习效率和泛化性能[18].小车的目标是爬到右边的山顶.小车在爬到山顶之前不会得到正反馈,因此小车对所处的环境是未知的.

由于小车的动力不足,不能直接爬到右边的山顶,因此它必须先向左爬,以获得充分的动能才能到达右边的山顶.小车的动力学方程如下:

$$\begin{cases} x_{t+1} = x_t + v_{t+1} \\ v_{t+1} = v_t + 0.001a_t - 0.0025\cos(3x_t) \end{cases} \tag{14}$$

其中,x 代表小车的位置,$x \in [-1.2, 0.5]$;v 代表小车的速度,$v \in [-0.07, 0.07]$;a_t 代表动作,动作空间 $A(s) \in \{-1, 0, 1\}$.当 $x_t = -1.2$ 时,小车速度为 0;当 $x = 0.5$ 时,小车目标完成.小车的起始点为 $x = -0.5$,$v = 0.0$.

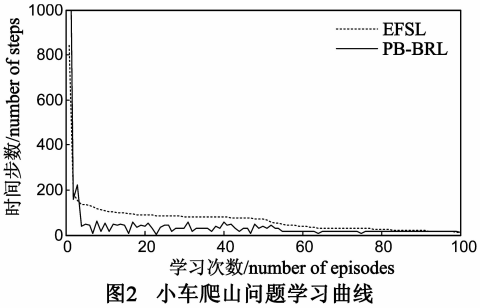

图2　小车爬山问题学习曲线

小车爬山学习控制系统的主要评价指标为小车从起始点位置爬到目标位置的时间步数(number of steps)和达到稳态所需要的学习次数(number of episodes).学习次数越少,算法学习过程过程越短,算法收敛性越好.时间步数越小,算法的实时性能就越好.为了测试本文所提 PB-BRL 学习算法的有效性,将本文算法与 EFSL(Enhanced Fuzzy Sarsa Learning)算法[18]进行对比,实验结果如图 2 所示.本文实验中,最大时间步数为 1000,学习率为 0.1,折扣因子为 0.9,温度参数的最大值和最小值分别为 0.1 和 0.001,采样周期为 0.02s.从图 2 可知,PB-BRL 算法经过大约 8 次学习后就能够在 15 至 20 个时间步数内实现小车爬山的目标.而 EFSL 经过大约 10 次学习后,需要大约 100 个时间步数内实现小车爬山的目标.如果 EFSL 要在 20 个时间步数实现小车目标,则至少需要学习 75 次.从实验结果可知,PB-BRL 比 EFSL 具有更好的收敛性和实时性.

5　总结

针对基于模型的贝叶斯强化学习中学习参数"维数灾"问题,本文采用可分解方法对未知学习参数进行降维.针对模型不确定环境下的动态系统,采用同时学习 DBNs 结构和未知参数的方法,可以有效地实现对模型不确定环境下动态系统的真实建模,从而解决应用建模难问题.将未知参数看成 MDPs 的隐变量,将 MDPs 的学习问题转化为 POMDPs 的规划问题.经过 MDPs 到 POMDPs 的转化,现有的 POMDPs 规划方法都可以应用到 MDPs 的学习中,解决了强化学习泛化难题.本文最后提出一种基于点的在线值迭代算法实现在线的规划和学习.实验结果表明 PB-BRL 算法可在较短时间内得到近似最大报酬值,为在线求解大规模贝叶斯强化学习提供可能.

参考文献

[1] 徐昕,沈栋,高岩青,等.基于马氏决策过程模型的动态系统学习控制:研究前沿与展望[J].自动化学报,2012,38(5):673 – 687.
Xu Xin, Shen Dong, Gao Yan-qing, et al. Learning control of dynamical systems based on Markov decision processes: research frontiers and outlooks [J]. Acta Automatica Sinica, 2012, 38(5): 673 – 687. (in Chinese)

[2] 刘海涛,洪炳熔,朴松昊,等.不确定性环境下基于进化算法的强化学习[J].电子学报,2006,34(7):1356 – 1360.
LIU Hai-tao, HONG Bing-rong, PIAO Song-hao, et al. Evolutionary algorithm based reinforcement learning in the uncertain environments[J]. Acta Electronica Sinica, 2006, 34(7): 1356 – 1360. (in Chinese)

[3] 刘全,李瑾,傅启明,等.一种最大集合期望损失的多目标 Sarsa(λ)算法[J].电子学报,2013,41(8):1469 – 1473.
LIU Quan, LI Jin, FU Qi-ming, et al. A multiple-goal Sarsa(λ) algorithm based on lost reward of greatest mass[J]. Acta Electronica Sinica, 2013, 41(8): 1469 – 1473. (in Chinese)

[4] Ross S, Pineau J, Chaib-draa B, et al. A Bayesian approach for

learning and planning in partially observable Markov decision processes[J]. Journal of Machine Learning Research, 2011, 12(1): 1729 – 1770.

[5] 高阳,胡景凯,王本年,等.基于 CMAC 网络强化学习的电梯群控调度[J].电子学报,2007,35(2):362 – 365.
GAO Yang, HU Jing-kai, WANG Ben-nian, et al. Elevator group control using reinforcement learning with CMAC[J]. Acta Electronica Sinica, 2007, 35(2): 362 – 365. (in Chinese)

[6] Doshi-VelezF, Pineau J, Roy N. Reinforcement learning with limited reinforcement: Using Bayes risk for active learning in POMDPs[J]. Artificial Intelligence, 2012, 187 – 188(1): 115 – 132.

[7] Poupart P, Vlassis N. Model-based Bayesian reinforcement learning in partially observable domains[A]. Proceedings of the International Joint Conference on Autonomous Agents and Multi Agent Systems[C]. New York: ACM Press, 2008. 1025 – 1032.

[8] Ross S, Pineau J. Model-based Bayesian reinforcement learning in large structured domains[A]. Proceedings of the 24th conference annual conference on uncertainty in artificial intelligence[C]. Cambridge, MA: AUAI Press, 2008. 476 – 483.

[9] Poupart P, Vlassis N, Hoey J, et al. An analytic solution to discrete Bayesian reinforcement learning[A]. Proceedings of the 23rd international conference on Machine learning[C]. New York: ACM Press, 2006. 697 – 704.

[10] Duff M. Optimal learning: Computational procedures for Bayes-adaptive Markov decision processes[D]. USA: University of Massassachusetts Amherst, 2002.

[11] Kearns M J, Koller D. Efficient reinforcement learning in factored MDPs[A]. Proceedings of the 16th International Joint Conference on Artificial Intelligence[C]. San Francisco: Morgan Kaufmann, 1999. 740 – 747.

[12] Guestrin C, Koller D, Parr R, et al. Efficient solution algorithms for factored MDPs[J]. Journal of Artificial Intelligence Research, 2003, 19(1): 399 – 468.

[13] Porta J M, Vlassis N A, Spaaan M T J, et al. Point-based value iteration for continuous POMDPs[J]. Journal of Machine Learning Research, 2006, 7(1): 2329 – 2367.

[14] 王雪松,张依阳,程玉虎.基于高斯过程分类器的连续空间强化学习[J].电子学报,2009,37(6):1153 – 1158.
WANG Xue-song, ZHANG Yi-yang, CHENG Yu-hu. Reinforcementlearning for continuous spaces based on gaussian process classifier[J]. Acta Electronica Sinica, 2009, 37(6): 1153 – 1158. (in Chinese)

[15] 仵博,吴敏,佘锦华.基于点的 POMDPs 在线值迭代算法[J].软件学报,2013,24(1):25 – 36.
Wu B, Wu M, She JH. Point-based online value iteration algorithm for POMDPs[J]. Journal of Software, 2013, 24(1): 25 – 36. (in Chinese)

[16] Wang Y, Won K S, Hsu D, et al. Monte Carlo Bayesian reinforcement learning[A]. Proceedings of the 29th International Conference on Machine Learning[C]. Edinburgh Scotland: Omni Press, 2012. 1135 – 1142.

[17] 刘春阳,谭应清,柳长安,等.多智能体强化学习在足球机器人中的研究与应用[J].电子学报,2010,38(8):1958 – 1962.
LIU Chun-yang, TAN Ying-qing, LIU Chang-an, et al. Application of multi-agent reinforcement learning in robot soccer[J]. Acta Electronica Sinica, 2010, 38(8): 1958 – 1962. (in Chinese)

[18] Vali D, Vahid J M, Majid N A. Exploration and exploitation balance management in fuzzy reinforcement learning[J]. Fuzzy Sets and Systems, 2010, 161(4): 578 – 595.

作者简介

仵 博 男,1979 年 12 月出生,河南桐柏人.2000 年、2003 年和 2013 年分别获得中南大学计算机科学与技术专业学士、硕士和博士学位.现为深圳职业技术学院副教授,从事序贯决策、机器学习和大数据方面的有关研究.
E-mail: wubo@szpt.edu.cn

郑红燕 女,1983 年 6 月出生,河南洛阳人.2004 年和 2007 年分别获得中南大学信息科学与工程学院学士和硕士学位.现为深圳职业技术学院高级工程师,从事模糊控制、智能决策和无线电认知方面的有关研究.
E-mail: zhenghongyan@szpt.edu.cn