贝叶斯学习与强化学习结合技术的研究*)

陈 飞 王本年 高 阳 陈兆乾 陈世福
(南京大学计算机软件新技术国家重点实验室 南京 210093)

摘 要 强化学习的研究需要解决的重要难点之一是:探索未知的动作和采用已知的最优动作之间的平衡。贝叶斯学习是一种基于已知的概率分布和观察到的数据进行推理,做出最优决策的概率手段。因此,把强化学习和贝叶斯学习相结合,使 Agent 可以根据已有的经验和新学到的知识来选择采用何种策略:探索未知的动作还是采用已知的最优动作。本文分别介绍了单 Agent 贝叶斯强化学习方法和多 Agent 贝叶斯强化学习方法:单 Agent 贝叶斯强化学习包括贝叶斯 Q 学习、贝叶斯模型学习以及贝叶斯动态规划等;多 Agent 贝叶斯强化学习包括贝叶斯模仿模型、贝叶斯协同方法以及在不确定下联合形成的贝叶斯学习等。最后,提出了贝叶斯在强化学习中进一步需要解决的问题。

关键词 贝叶斯学习,强化学习,单 Agent,多 Agent

Research on the Combihation of Bayesian Learning and Reinforcement Learning

CHEN Fei WANG Ben-Nian GAO Yang CHEN Zhao-Qian CHEN Shi-Fu
(State Key Lab. for Novel Software Technology, Department of Science and Technology, Nanjing University, Nanjing 210093)

Abstract A central problem in reinforcement learning is balancing exploration of untested actions against exploitation of actions that are known to be good. Bayesian learning is a probability method that makes optimal decision based on known probability distribution and recently observed data. So combination of Bayesian learning and reinforcement learning the agent can choose the strategy of exploration or exploitation based on its own experience and newly incoming knowledge. In this paper, we introduce single-agent Bayesian reinforcement learning and multi-agent Bayesian reinforcement learning. Single-agent Bayesian reinforcement learning includes Bayesian Q-learning, model-based Bayesian learning and Bayesian DP, and multi-agent Bayesian reinforcement learning includes Bayesian imitation, Bayesian coordination and Bayesian reinforcement learning for coalition formation under uncertainty. At last, some unsolved problems in Bayesian reinforcement learning are discussed.

Keywords Bayesian learning, Reinforcement learning, Single-agent, Multi-agent

1 引言

强化学习由 Minsky 在 20 世纪 50 年代首次提出,目前已经成为机器学习领域的研究热点之一[1]。强化学习是指从环境状态到行为映射的学习,以使系统行为从环境中获得的累积奖赏值最大,通过试错(trial-and-error)的方法来发现最优行为策略[2]。其中需要解决的重要难点之一是:开采和探索的平衡(exploration-exploitation tradeoff),就是选择采用已知的最优动作还是探索未知的动作[2,3]。贝叶斯学习提供了推理的一种概率手段,它基于假定待考查的量遵循某种概率分布,根据该概率和观察到的数据进行推理,以做出最优的决策[4]。因此,把强化学习和贝叶斯学习相结合,为环境模型建立一系列可能的概率分布,通过概率的方法来估计环境的动态变化,使 Agent 可以基于已有的经验和新学到的知识来选择采用何种策略:探索未知环境还是采用已知的"最优"的动作。这样可以很好地解决开采和探索的平衡问题。

本文分别介绍了单 Agent 贝叶斯强化学习方法和多 Agent 贝叶斯强化学习方法。在单 Agent 强化学习方法中,首先介绍贝叶斯 Q 学习[5],接着介绍贝叶斯模型学习[6],最后是贝叶斯动态规划方法[7];在多 Agent 强化学习方法中,依次介绍贝叶斯模仿模型[8]、贝叶斯协同方法[9]以及在不确定下联合形成中的贝叶斯学习[10]。最后,提出了在贝叶斯强化学习研究中进一步需要解决的问题。

2 单 Agent 贝叶斯强化学习

传统的强化学习可以看成一个单 Agent 系统[11],下面介绍单 Agent 贝叶斯强化学习的几种应用方法。这里假设学习环境是马尔可夫型的,则顺序型强化学习问题可以通过马尔可夫决策过程(Markov Decision Process, MDP)建模。MDP 是一个四元组$\langle S, A, p_t, p_r\rangle$,其中 S 为环境状态集合;A 为行为集合;$p_t(s \xrightarrow{a} t)$是状态转换模型,表示 Agent 在状态 s 下执行动作 a 转化到状态 t 的概率;$p_r(r|s,a)$是奖赏模型,表示 Agent 在状态 s 下执行动作 a 得到奖赏 r 的概率。

2.1 贝叶斯 Q-学习

众所周知,Q-学习是目前最具代表性的一种模型无关单 Agent 强化学习算法[12,13]。Q-学习算法的实现原理简单,目的是求出最大折扣回报的动作,基本形式如式(1)、(2):

*)本课题得到国家自然科学基金(60475026)、国家"973"重点基础研究发展计划基金项目(2002CB312002)和江苏省自然科学基金(BK2004079)的资助。陈 飞 硕士研究生,研究领域:机器学习、数据挖掘、神经网络;陈世福 教授,博士生导师,研究领域:人工智能、机器学习。

$$Q(s,a)=E[R(s,a)]+\gamma\sum_{s'}T(s,a,s')\max_{a'}Q(s',a') \quad (1)$$

$$Q(s,a)=[1-a]Q(s_t,a_t)+a(r+\gamma\max_a Q(s_{t+1},a)) \quad (2)$$

式(1)中 $Q(s,a)$ 表示 Agent 在状态 s 下采用动作 a 所获得的最优奖赏折扣和，最优策略为在状态 s 下选取 Q 值最大的行为。Q-学习首先初始化 Q 值；然后 Agent 在状态 s_t 下采用动作 a_t 得到奖赏 r，依据式(2)更新 Q 值；如此迭代循环，直至学习过程结束。Watkins 等人利用随机过程和不动点理论，证明了当 a 满足一定条件时 MDP 模型 Q-学习过程的收敛性[14]。

在 Q-学习中引进贝叶斯方法，用概率分布表示 Agent 对每个状态动作对的 Q 值估计的未知度。通过保存和传播 Q 值的分布，而不是点估计，可以做出更可靠的决定选择下一步动作。该方法的前提假设是：假设在任意状态 s 下的动作 a 获得的总折扣奖赏满足标准概率分布。

Q 值的未知度被表示为一定的概率分布。通过引入对于 Q 值的未知度，把无模型贝叶斯强化学习建立在 Q 值概率分布这个基础上。由于在学习过程中并不知道环境模型，因此对每个 (s,a) 都有一个 Q 值分布。处理探索和开采两者平衡时，基于一个著名的信息决策思想：信息价值，就是通过探索获得的信息使得未来决策质量的期望改进[15]。采用了称为 Myopic-VIP(Myopic value of perfect information)的新方法，可以直接评估探索和开采权衡，通过比较探索可获得的预期收益(改进策略)与采用已知最优动作可获得的预期回报来选择策略。

新知识分两种情况改变 Agent 策略：(a)新知识表明原来被认为的最优动作就是最佳选择；(b)新知识表明原来被认为最佳的动作其实比别的动作差。

定义 $q_{s,a}$ 是 MDP 中 $Q^*(s,a)$ 的可能取值，$Q^*_{s,a}$ 是 $q_{s,a}$ 的真实值，学习 $q^*_{s,a}$ 可以获得什么，也就是新知识如何改变 Agent 的未来奖赏值。分为下面两种情况考虑：

情况(a)时，假设 a_1 是最佳动作，即对于所有其他动作 a'，$E[q_{s,a_1}]\geqslant E[q_{s,a'}]$。然而新知识表明 a 是更好的动作，即 $q^*_{s,a}>E[q_{s,a_1}]$，则定义 Agent 执行 a 而不是 a^* 时，获得价值增益是 $q^*_{s,a}-E[q_{s,a_1}]$。

情况(b)时，假设 a_1 是最佳动作，a_2 是第二最佳动作。如果新知识表明 $q_{s,a_1}<E[q_{s,a_2}]$，那么 Agent 应该执行动作 a_2 而不是 a_1，获得价值增益是 $E[q_{s,a_2}]-q^*_{s,a_1}$。

这样，学习 $q^*_{s,a}$ 获得的价值增益分为式(3)所示的几种情况进行计算。

$$\mathrm{Gain}_{s,a}(q^*_{s,a})=\begin{cases}E[q_{s,a_2}]-q^*_{s,a}\\ (\text{if } a=a_1 \text{ and } q^*_{s,a}<E[q_{s,a_2}])\\ q^*_{s,a}-E[q_{s,a_1}]\\ (\text{if } a\neq a_1 \text{ and } q^*_{s,a}>E[q_{s,a_2}])\\ 0\\ (\text{otherwise})\end{cases} \quad (3)$$

在式(3)中，a_1 是预计最大价值的动作，a_2 是预计第二价值的动作。

$q_{s,a}$ 的理想信息的预期价值，求法如式(4)所示。

$$\mathrm{VPI}(s,a)=\int_{-\infty}^{\infty}\mathrm{Gain}_{s,a}(x)\Pr(q_{s,a}=x)\mathrm{d}x \quad (4)$$

我们选择的动作是使得式(5)取最大值的动作。

$$\mathrm{VPI}(s,a)-(\max_{a'}E[q_{s,a'}]-E[q_{s,a}]) \quad (5)$$

明显看出，该策略选择的动作其实是使得式(6)取最大值的动作：

$$E[q_{s,a}]+\mathrm{VPI}(s,a) \quad (6)$$

贝叶斯最优动作选择方法存在的难点是要计算一个动作序列上的积分，而该积分被证明是难以被直接计算的[16]。因此，必须采取估计和逼近的方法。每次执行一次动作和状态的转换后，可采用简单更新[17]或者混合更新 Q 值的估计分布的方法。

混合更新的方法先定义 $p(\mu_{s,a},\tau_{s,a}|R)$ 为在观察到折扣回报 R 后 $\mu_{s,a},\tau_{s,a}$ 的后验分布，其中 $\mu_{s,a}$ 为 R 概率分布的均值，$\tau_{s,a}$ 为 R 概率分布的精度(方差的倒数)。如果观察到的估计值 $R_t=x$，更新 $R_{s,a}$ 的后验概率是 $p(\mu_{s,a},\tau_{s,a}|r+\gamma x)$，则混合后验概率更新如式(7)所示：

$$p^{\mathrm{mx}}_{r,t}(\mu_{s,a},\tau_{s,a})=\int_{-\infty}^{\infty}p(\mu_{s,a},\tau_{s,a}|r+\gamma x)p(R_t=x)\mathrm{d}x \quad (7)$$

混合更新的试验效果比简单更新明显变好，但是还未从理论上证明其收敛性。

2.2 基于模型的贝叶斯探索

是否采用模型是强化学习中的一个争论热点。无模型的方法不需要精确估计周围环境的动态变化，直接学习近似最优策略；基于模型的方法则建立一个模型来描述环境动态变化，使用该模型来求解动作的预期价值。基于模型的方法的优点之一是 Agent 避免代价昂贵的重复步骤，通过模型中模拟的步骤进行学习[18]。

目前绝大多数已有的基于模型的方法都是用一些简单的估计方法来学习环境，对于环境的动态变化采用点估计。这种点估计的方法并不能反映 Agent 对环境动态变化各个方面的不确定性。因此可以把贝叶斯学习引入到基于模型的强化学习之中：基于一定的合理假设，通过一定的模型从先验经验中推出后验概率分布，并且在执行动作时不停地更新该模型的概率分布。

贝叶斯模型学习方法是对可能的 MDP 保持一个信任状态。置信区间 μ 定义了一个概率密度 $P(M|\mu)$。给定一个经验四元组 $\langle s,a,r,t\rangle$，可计算出后验置信状态 $\mu^\circ\langle s,a,r,t\rangle$，方法如式(8)所示：

$$\begin{aligned}&P(M|\mu^\circ\langle s,a,r,t\rangle)\\ &\propto P(\langle s,a,r,t\rangle|M)P(M|\mu)\\ &=P(s\xrightarrow{a}t|M)P(s\xrightarrow{a}r|M)P(M|\mu)\end{aligned} \quad (8)$$

计算时参数化 MDP，分为转换模型参数 $\theta^t_{s,a,t}$ 和奖赏模型参数 $\theta^r_{s,a,r}$。同时假设信任状态 $P(\theta|\mu)$ 满足参数相互独立，可理论证明出在该假设条件下后验概率 $P(\theta|\mu^\circ\langle s,a,r,t\rangle)$ 也满足参数相互独立。动作选取依然采用贝叶斯 Q-学习的原理，选择使式(6)最大化的动作。

采用以下 4 种估计 Q 值分布的方法：朴素取样法、重要取样法、修正全局取样法以及局部取样法。

(1)朴素取样法：是一种很简单的方法，但是需要有效的取样过程。得到样本后，估计 Q 值的均值，如式(9)所示：

$$E[q_{s,a}]\approx\frac{1}{\sum_i\omega^i_\mu}\sum_i\omega^i_\mu q^i_{s,a} \quad (9)$$

采用相似的办法估计 VPI 值，具体求法如式(10)所示：

$$\mathrm{VPI}(s,a)\approx\frac{1}{\sum_i\omega^i_\mu}\sum_i\omega^i_\mu\mathrm{Gain}_{s,a}(q^i_{s,a}) \quad (10)$$

(2)重要取样法：朴素取样法的主要问题是评价 Agent 的每个动作都需要通过好几个全局计算。但重要取样法重用相

同的取样 MDP 来避免重复的计算，调整取样的权值来纠正取样分布和目标分布的差异。具体计算如式(11)所示：

$$\omega_{\mu'}^{i}=\frac{\Pr(M^{i}|\mu')}{\Pr(M^{i}|\mu)}w_{\mu}^{i} \tag{11}$$

(3)修正全局取样法：明显看出重要取样法计算也是很复杂的，为了减少计算 MDP 的代价，可以不停更新每个取样 MDP。如果原来的样本 M^i 是从 $\Pr(M|\mu)$ 中抽样，观察到一个经验元组$\langle s,a,r,t\rangle$后，修正的 M^i 从 $\Pr(M|\mu^{\circ}\langle s,a,r,t\rangle)$中抽样。具体的做法与优先扫描相似，不同之处是修正全局取样法同时运行 k 个实例。

(4)局部取样法：与采用 MDP 的全局取样方法不同的是，为每一对(s,a)保存一个 Q 值分布估计，使用局部的贝尔曼(Bellman)更新传播的方法来更新这些概率分布。贝尔曼等式的基本形式如式(12)所示：

$$q_{s,a}=E[p_R(s\xrightarrow{a}r)]+\gamma\sum_{s'\in S}p_T(s\xrightarrow{a}s')\max_{a'}q_{s',a'} \tag{12}$$

从式(12)可看出，从 $q_{s,a}$ 中取样就可以计算出 Q 值，只要重复这种取样步骤 k 次，便得到 $q_{s,a}$ 的 k 个样本。

2.3 贝叶斯动态规划

Malcolm 等[7]提出一种贝叶斯动态规划方法，是对贝叶斯模型学习的改进：从模型中取样出一个假设，通过动态规划求出基于此假设的贪心算法。Dearden 等[6]已经提出了几种从模型中取样的估计方法，在这些取样方法中，尽可能多地使用重取样；而在 Malcolm 的方法中，在几个时间步骤中保留相同的一个假设，保持同一个面向目标的探索策略，而且不需要采用近似度量的方法(如 myopic VPI)。这样做，除了可以降低计算成本，还使得探索策略在某个时期内保持不变。其中的基本原理是基于人的正常学习过程，就是在新假设形成之前，可以通过一系列连贯的动作来测试某个假设的效果。这个方法使得每次试错时只需要建立一个模型假设，有利于解决比较庞大和复杂的问题。

动态规划是通过重复估计替换的方法来解决具有一定约束条件的问题[19]。使用动态规划的目的有两个：(1)利用极大近似的 MDP 各参数求出最优策略；(2)对于每个从 MDP 模型参数的后验概率分布中推出的假设，求出对应的最优策略。

首先为 $Q(s,a)$ 建立一个明确的 MDP 模型，然后在该 MDP 模型上进行动态规划，如式(13)所示。

$$Q(s,a)=E[R(s,a)]+\gamma\sum_{s'}T(s,a,s')\max_{a'}Q(s',a') \tag{13}$$

当回报分布和状态转移概率已知时，等式(13)产生一系列同步的非线性 Q 值等式(每个等式对应一个状态动作对)。动态规划的价值迭代方法直接用估计重复替换来处理这一系列等式。动态规划的优点是不需要使用学习率，缺点是必须假定回报分布和转换概率是统计不变的。因此，动态规划不适合动态变化显著的环境，在比较稳定的环境中效果很好。

假设转换矩阵是稀疏的，p 中只有一定数目的元素是非零，用一种分级的狄利特雷形式统计转换的未知性，采用高斯密度分布建立瞬间奖赏的模型。

定义在概率分布的均值 μ 上的先验概率密度为：

$$f(\mu)\propto N(\mu_0,\sigma^2) \tag{14}$$

这样，结合先验密度和观察到的证据，便可求出模型参数的后验概率分布。设 n 为观察到的事件(由参数 p 表示)发生次数，则通过贝叶斯公式求出 $f(p)$ 的后验概率，如式(15)所示：

$$P(p|n)=\frac{P(n|p)P(p)}{P(n)} \tag{15}$$

则求得在 μ 上的后验概率，如式(16)所示：

$$f(\mu)\propto N(\bar{x},\sigma^2/n) \tag{16}$$

在每次试验开始的时候或者学习中固定的时间间隔点，从 MDP 参数的后验分布中产生一个无偏的假设。从每个状态-动作对中产生 σ,μ 和 p 的值。其中，μ 是从高斯分布中取样求出；σ,p 是把均匀分布 $U[0,1]$映射到从适应性数字积分(adaptive numerical integration)得到的累积概率密度函数后求得。这样，便从 MDP 模型中产生了一个假设，再从该假设中选择最优的动作。

3 多 Agent 贝叶斯强化学习

本文主要从机器学习的角度考察多 Agent 环境[11]，讨论同构 Agent 可通信与异构 Agent 可通信两种情况下，将贝叶斯学习与强化学习相结合的方法。多 Agent 强化学习方法与传统的单 Agent 强化学习方法很相似，只是增加了学习复杂性[31,32]。因此，贝叶斯多 Agent 强化学习可解决与单 Agent 强化学习中存在的相似问题，加快多 Agent 学习速度，解决多 Agent 学习中的难点。

3.1 贝叶斯模仿模型

在多 Agent 环境中，通信可以使某个 Agent 学习到其他 Agent 的经验，加快学习的速度[20]。但是采用显式的通信手段有许多弊端：需要一个通信通道、一种可以充分表达含义的语言、一个沟通不同 Agent 的翻译器以及通信的动机等；且通信需要付出一定的代价。而基于观察的技术却不同，学习的 Agent 只需观察其他 Agent 的外在表现，比如模仿[21]和逆向强化学习方法[22]等。隐含模仿，就是通过被动观察来实现隐含通信[23,24]，与传统模仿模型不同之处在于：学习者不需要准确复制其他 Agent 的行为。把隐含模仿和贝叶斯框架结合起来，建立贝叶斯模仿模型；在本方法中，观察者直接把观察到的导师知识结合到自己环境的增量模型之中。

这里，假设各 Agent 的行为能力与目的相似。定义导师控制的 MDP 为$\langle S,A_m,R_m,D\rangle$，观察者控制的 MDP 为$\langle S,A_o,R_o,D\rangle$，两者有相同的状态空间和动态变化情况。观察者通过自己的动作经验和对导师的观察，更新自己对于 D 的信任概率，如式(17)所示：

$$\begin{aligned}&P(D|H_o,A_o,H_m)\\&=aPr(H_o,H_m|D,A_o)P(D)\\&=aPr(H_o|D,A_o)Pr(H_m|D)P(D)\end{aligned} \tag{17}$$

对于观察者而言，$P(D^{s,a})$的更新分为两种情况：

(1)当导师的策略 π_m^s 和观察者的动作 a 不一致时，更新时不用考虑导师的影响，如式(18)所示：

$$P(D^{s,a}|H_o^{s,a})=aPr(H_o^{s,a}|D^{s,a})P(D^{s,a}) \tag{18}$$

(2)当导师的策略 π_m^s 和观察者的动作 a 一致时，更新要加入导师的经验，如式(19)所示：

$$\begin{aligned}&P(D^{s,a}|H_o^{s,a},H_m^s,\pi_m^s=a)\\&=a\Pr(H_o^{s,a},H_m^s|D^{s,a},\pi_m^s=a)P(S^{s,a}|\pi_m^s=a)\\&=a\Pr(H_o^{s,a}|D^{s,a})\Pr(H_m^s|D^{s,a},\pi_m^s=a)P(D^{s,a})\end{aligned} \tag{19}$$

定义 $n^{s,a}$ 为 $P(D^{s,a})$ 的先验参数向量，$c_o^{s,a}$ 是观察者从在状态 s 执行动作 a 的计数，c_m^s 是导师从状态 s 转换的计数，后验增量模型满足狄利特雷分布，则观察者和导师的计数的更新公式如式(20)所示：

$$P(D^{s,a}\mid H_o^{s,a},H_m^s,\pi_m(s)=a)$$
$$=P(D^{s,a};n^{s,a}+c_o^{s,a}+C_m^s) \tag{20}$$

观察者并不知道导师的动作属于上述哪种情况，因此结合两种情况的环境更新公式如式(21)所示：

$$P(D^{s,a}\mid H_o^{s,a},H_m^s)$$
$$=\Pr(\pi_m^s=a\mid H_o^{s,a},H_m^s)P(D^{s,a};n^{s,a}+c_o^{s,a}+c_m^s)$$
$$+\Pr(\pi_m^s\neq a\mid H_o^{s,a},H_m^s)P(D^{s,a};n^{s,a}+c_o^{s,a}) \tag{21}$$

最后要解决的问题是，观察者对于导师策略的信任概率的更新方法如式(22)所示：

$$\Pr(\pi_m\mid H_m,H_o)$$
$$=\alpha\Pr(H_m\mid\pi_m,H_o)\Pr(\pi_m\mid H_o)$$
$$=\alpha\Pr(\pi_m)\int_{D\in d}\Pr(H_m\mid\pi_m,D)P(D\mid H_o) \tag{22}$$

贝叶斯模仿者的学习过程如下：每一步中，Agent 观察自己和导师的状态变换，使用上述的方法更新自己的模型概率，再用有效的方法更新 Agent 的效用函数，随后基于更新的效用函数，选择执行合适的动作。如此重复迭代循环。

3.2 贝叶斯协同方法

多 Agent 强化学习算法(MARL)要求保证能(最终)收敛到正确的均衡，通常是通过各 Agent 协调最终达到各自的均衡。目前研究表明，可以采用一系列启发式搜索达到最佳均衡[25,26]，当与传统的强化学习一样，搜索足够的状态空间才能保证收敛，因此需要考虑达到收敛的代价问题[27]。MARL 也存在开采和探索的平衡问题，而且有新的特点：某 Agent 的动作选择也影响到其他 Agent 对未来动作的选择。因此，引入贝叶斯协同模型，解决 MARL 中的广义开采-探索平衡的问题。具体做法是把 Dearden 等[6]提出的贝叶斯模型推广到 MARL 中。

定义一个贝叶斯 MARL Agent(BA)，有一些可能模型空间的先验分布和可供其他 Agent 利用的策略空间。BA 信任状态的形式是 $b\langle P_M,P_S,s,h\rangle$，其中：$P_M$ 是已知模型空间的密度；P_S 是其他智能体采用的可能策略的一个联合密度；s 是系统当前的状态；h 是游戏当前历史各相关方面的总和，足够预测任何与 P_S 一致的 Agent 做出的动作选择。更新方法如式(23)所示：

$$b'=b(\langle s,a,r,t\rangle)=\langle P'_M,P'_S,t,h'\rangle \tag{23}$$

其中

$$P'_M(m)=z\Pr(t,r\mid a,m)P_M(m)$$
$$P'_S(\sigma_{-i})=z\Pr(a_{-i}\mid s,h,\sigma_{-i})P_S(\sigma_{-i})P_S(\sigma_{-i}) \tag{24}$$

在式(24)中，策略轮廓 σ 是一个选择动作的策略的集合，与每个 Agent 一一对应。σ_i 表示 σ 中 Agent i 的部分，σ_{-i} 表示一个简化的策略轮廓，该轮廓包括除 i 的策略之外的所有策略。用 $\sigma_{-i}\circ\sigma_i$ 表示在 σ_{-i} 中增加 σ_i 后共同组成的(完整的)轮廓。

因此，重新定义贝叶斯探索模型。在信任状态 b 时动作 a_i 的效用分为两个部分：关于当前信任状态的期望效用与它对当前信任状态的影响，就是动作的信息期望效用(EVOI)。EVOI 是不可直接求出来的，但可通过在信任状态 MDP 上的 Bellman 等式和"对象层次"的期望效用结合求出，如式(25)所示：

$$Q(a_i,b)=\sum_{a_i}\Pr(a_{-i}\mid b)\sum_t\Pr(t\mid a_i\circ a_{-i},b)$$
$$\sum_r\Pr(r\mid a_i\circ a_{-i},b)[r+\gamma V(b(\langle s,a,r,t\rangle))]$$
$$V(b)=\max_{a_i}Q(a_i,b) \tag{25}$$

该等式可解决 POMDP 和高维连续 MDP 问题。在实际应用中，需要一系列计算技巧和逼近方法，具体定义如式(26)、(27)、(28)所示：

$$\Pr(a_{-i}\mid b)=\int_{\sigma-i}\Pr(a_{-i}\mid\sigma_{-i})P_S(\sigma_{-i}) \tag{26}$$

$$\Pr(t\mid a,b)=\int_m\Pr(t\mid s,a,m)P_M(m) \tag{27}$$

$$\Pr(r\mid b)=\int_m\Pr(r\mid s,m)P_M(m) \tag{28}$$

采用 myopic 方法选择动作：给定信任状态 b，每个动作 $a_i\in A$ 的 myopic Q 函数定义如式(29)、(30)所示：

$$Q_m(a_i,b)=\sum_{a_{-i}}\Pr(a_{-i}\mid b)\sum_t\Pr(t\mid a_i\circ a_{-i},b)$$
$$\sum_r\Pr(r\mid a_i\circ a_{-i},b)[r+\gamma V_m(b(s,a,r,t))] \tag{29}$$

$$V_m(b)=\max_{a_i}\int_m\int_{\sigma_{-i}}Q(a_i,s\mid m,\sigma_{-i})P_M(m)P_S(\sigma_{-i}) \tag{30}$$

选择动作时采用 2.1 节中贝叶斯 Q-学习所使用的方法，采用式(3)、(4)、(5)计算。

3.3 在不确定中联合形成的贝叶斯强化学习

联合形成是对策论的研究热点[28]，在 MARL 中得到日益关注，用于动态组成 Agent 的合作团队[29,30]。联合形成研究通常假定潜在的联合价值是已知的，并且很少涉及当 Agent 缺少对潜在合作伙伴能力了解的足够知识的情况。除去这些不切合实际的假设，Chalkiadakis 等[10]建立了一个采用贝叶斯多 Agent 强化学习方法的模型，减少编队的成员关于编队价值和其他成员的能力的不确定性。同时，还引进了与该模型相适应的贝叶斯核心(BC)，一种在不确定情况下联合信息的稳定性概念，描述了可收敛到 BC 的动态联合[30]形成过程。在该模型中，每个 Agent 保留了对其他 Agent 类属的准确信念，并且选择动作和联合时既考虑直接效用又考虑它们的信息价值[15]。

假设 $N=\{1,\cdots,n\}(n>2)$ 是一个选手集合，子集 $S\subseteq N$ 被称为一个联合，属于同一个联合的 Agent 会为了共同的利益协调各自行为。联合结构是 Agent 集合的划分，包括穷尽的联合和分离的联合。联合形成是各 Agent 形成这种联合的过程，通常是为了解决某个问题而协调大家的行为。

贝叶斯联合形成问题包括 6 个主要组成部分：Agent 集合、类型集合、联合动作集合、结果或状态集合、奖赏函数以及 Agent 对类型的信任度。

贝叶斯核(简称 BC)定义：一对联合结构和要求向量 $\langle CS,\vec{d}\rangle$，$C_i$ 表示成员 i 的 $C\in CS$。则在一个贝叶斯联合问题中，$\langle CS,\vec{d}\rangle$ 属于贝叶斯核，如果满足下列条件：对所有的 $C\in CS$，存在某个 $\alpha\in A_C$，使得没有 $S\in N$ 存在某个动作 $\beta\in A_S$ 和向量 $\vec{d}_S$，使得 $\bar{p}^{i}_{i,S}(\beta,\vec{d}_S)>\bar{p}^{i}_{i}(\alpha,C_i)$，$\forall i\in S$。

采用贝叶斯方法迭代进行联合形成，联合中的每个成员不停地更新对它合作伙伴种类的信任度。具体的方法如式(30)所示：

$$B_i^{t+1}(\vec{t_C})=z\Pr(s\mid\alpha,\vec{t_C})B_i^t(\vec{t_C}) \tag{30}$$

这里 z 是一个标准化的常量。

可以把此处的最佳学习问题看成部分感知马尔可夫问题(POMDP)，或者信念状态马尔可夫问题。于是，把标准的 Bellman 等式修改为式(31)、(32)的形式。

$$Q_i(C,\alpha,\vec{d_C},B_i)$$
$$=\sum_s\Pr(s\mid C,\alpha,B_i)[r_iR(s)+\gamma V_i(B_i^{s,a})]$$
$$=\sum_{\vec{t_C}}B_i(\vec{t_C})\sum_s\Pr(s\mid C,\alpha,\vec{t_C})[r_iR(s)+\gamma V_i(B_i^{s,a})] \tag{31}$$

$$V_i(B_i)=\sum_{C|i\in C,\vec{d_C}}\Pr(C,a,\vec{d_C}\,|\,B_i)Q_i(c,a,\vec{dc},B_i) \qquad (32)$$

接着，使用4种强化学习方法：非近视完全协商(NM-FN)，Agent在联合形成时进行完全协商；近视完全协商(M-FN)，Agent进行完全协商来决定每一步的联合；近视单步建议(M-OSP)，近视的Agent使用信任度来估计联合价值，但在形成联合过程中不进行完全协商；非近视单步建议(NM-OSP)，是NM-FN和M-OSP的结合。完全协商的优点是在强化学习每个阶段之后，联合的结构处在一个稳定状态，并且Agent可以在协商中更新对其他Agent的信任度。单步建议方法的优点是使Agent更方便了解结构空间，在充分考虑"实时"性能的情况下效果最好(因为在每个学习阶段后无法有足够长的时间进行协商)。

总结和展望 把贝叶斯引入强化学习之中，可以很好地解决强化学习的重要难点之一：探索和开采的平衡。本文分别介绍了单Agent贝叶斯强化学习方法和多Agent贝叶斯强化学习方法。理论分析和试验结果表明，贝叶斯强化学习方法可以明显提高强化学习的性能、速度以及加快算法的收敛等。

目前，贝叶斯学习与强化学习结合技术已得到广泛应用，取得了显著的效果，但仍有许多问题需进一步研究，主要包括：(1)把贝叶斯强化学习方法应用到实际问题和具体环境中，同时要提出更好的贝叶斯积分的函数逼近方法；(2)考虑如何处理比较复杂的问题(比如连续值状态)、更复杂的模型表示以及很大的状态空间等；(3)考虑把贝叶斯强化学习应用到部分感知马尔可夫环境(POMDP)研究中；(4)考虑可否把分层思想引入贝叶斯强化学习方法中；(5)把单Agent贝叶斯学习方法推广到多Agent系统中；(6)把贝叶斯强化学习方法应用到多Agent学习的更多领域中。

参 考 文 献

1 Sutton R S, Barto S. Reinforcement learning. Cambridge, MA: MIT Press, 1998

2 Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey. J Artificial Intelligence Research, 1996, 4: 237～285

3 Gao Y, Chen SF, Lu X. Research on reinforcement learning technology: a review. Acta Automatica Sinica, 2004, 30(1): 76～90(in Chinese with English abstract)

4 Mitchell T M. Machine learning. Columbus, OH: McGraw-Hill, 1997

5 Dearden R, Friedman N, Russell S. Bayesian Q-learning. In: Proc. s of the Fifteenth National Conference on Artificial Intelligence (AAAI-98), Menlo Park, CA: AAAI Press

6 Dearden R, Friedman N, Andre D. Model based Bayesian Exploration. In: Proceedings of Fifteenth Conference on Uncertainty in Artificial Intelligence. San Francisco, Morgan Kaufmann, 1999. 150～159

7 Strens M. A Bayesian Framework for Reinforcement Learning. In: Proc. of the Seventeenth International Conference on Machine Learning (ICML-2000), Stanford University, California, June 29-July 2, 2000

8 Price B, Boutilier C. A Bayesian Approach to Imitation in Reinforcement Learning. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-2003), 2003. 712～720. Chalkiadakis G, Boutilier C. Coordination in multiagent reinforcement learning: a Bayesian approach. AAMAS, 2003. 709～716

9 Chalkiadakis G, Boutilier C. Coordination in Multiagent Reinforcement Learning: A Bayesian Approach. In: International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'03), 2003. 709～716

10 Chalkiadakis G, Boutilier C. Bayesian Reinforcement Learning for Coalition Formation under Uncertainty. In: International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), 2004. 1090～1097

11 Stone P, Veloso M. Multiagent Systems: A survey from a Machine Learning Perspective. Autonomous Robots, 2000, 8: 345～383

12 Watkins C J C H. Learning from delayed rewards: [Ph D Thesis]. Psychology Department, Cambridge University, Cambridge, U K, 1989

13 Watkins C J, Dayan P. Q-learning. Machine Learning, 1992, 8(3): 279～292

14 Tsitsiklis J N. Asynchronous Stochastic Approximation and Q-learning. Machine Learning, 1994, 16(3): 185～202

15 Howard R A. Information value theory. In: IEEE Transactions on Systems Science and Cybernetics SSC-2, 1966. 22～26

16 Martin J J. Bayesian decision problems and Markov chains. New York: John Wiley, 1967

17 Wyatt J. Exploration and Inference in Learning from Reinforcement: [PhD thesis]. Department of Artificial Intelligence, University of Edinburgh, 1997

18 Sutton R S. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In: Proc. of the Seventh Int. Conf. on Machine Learning, Morgan Kaufmann, 1990. 216～224

19 Bellman R E. Dynamic programming. Princeton, NJ: Princeton University Press, 1957

20 Sycara K. Negotiation planning: An AI approach. European Journal of Operational Research, 1990, 46: 216～234

21 Mataric M J. Visuo-motor primitives as a basis for learning by imitation: Linking perception to action and biology to robotics. In Imitation in Animals and Artifacts. Cambridge, MA: MIT Press, 2004. 392～422

22 Ng A Y, Russell S. Algorithms for inverse reinforcement learning. In: Proc. Seventeenth Intl Conf on Machine Learning, Morgan Kaufmann, San Francisco, CA, 2000. 663～670

23 Price B, Boutilier C. Implicit imitation in multiagent reinforcement learning. In: Proc. Sixteenth International Conference on Machine Learning, Bled, SI, 1999. 325～334

24 Price B, Boutilier C. Imitation and reinforcement learning in agents with heterogeneous actions. In: Proc. Fourteenth Canadian Conf on Artificial Intelligence Ottawa, 2001. 11～120

25 Lauer M, Riedmiller M. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In Proceedings of the Seventeenth International Conference on Machine Learning, Stanford, CA, 2000. 535～542

26 Wang X, Sandholm T. Reinforcement learning to play an optimal nash equilibrium in team markov games. Advances in Neural Information Processing Systems 15 (NIPS-2002), Vancouver, 2002

27 Boutilier C. Sequential optimality and coordination in multiagent systems. In: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, Stockholm, 1999. 478～485

28 Myerson R. Game Theory: Analysis of Conflict. Harvad University Press. 1991

29 Bancerjee B, Sen S. Selecting Partners. In: Sierra C, Gini M, Rosenschein J, eds. Proceedings of the Fourth International Conference on Autonomous Agents, Barcelona, Catalonia, Spain, ACM Press, 2000. 261～262

30 Konishi H, Ray D. Coalition Formation as a Dynamic Process. Boston College Working Papers in Economics 478, 2002

31 Littman M L. Markov games as a framework for multi-agent reinforcement learning. In: Proc. Eleventh Intl Conf on Machine Learning, New Brunswick, NJ, 1994. 157～163

32 Hu J, Wellman M P. Multiagent reinforcement learning: theoretical framework and an algorithm. In: Proc. Fifteenth Intl Conf. on Machine Learning, Madison, Wisconsin, 1998. 242～250