View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Xiamen University Institutional Repository

一种交互式动态影响图的改进算法

李 波 罗 键 尹华一 田 乐

(厦门大学 信息科学与技术学院 厦门 361005)

摘 要 交互式动态影响图(I-DIDs)是基于概率图形理论的多智能体动态交互决策的图模型. 为缓解该模型状态空间随时间片增加呈指数级增长的趋势,文中基于行为等价的基本思想压缩状态空间,提出构建 Epsilon 行为等价类的方法: 利用有向无环图表示其它 Agent 可能的信度和行为,把信度在空间上接近的模型聚为一类,实现自顶向下合并行为等价模型. 该过程避免求解状态空间中的所有候选模型,节省了存储空间和计算时间. 模型实例上的仿真结果显示了该算法的有效性.

关键词 Agent 建模,交互式动态影响图,动态决策,ε-行为等价,信度-行为图

中图法分类号 TP 181

An Improved Algorithm for Interactive Dynamic Influence Diagrams

LI Bo, LUO Jian, YIN Hua-Yi, TIAN Le

(*Department of Information Science and Technology, Xiamen University, Xiamen* 361005)

ABSTRACT

Interactive Dynamic Influence Diagrams (I-DIDs), as graphic models based on probabilistic graphical theory, are proposed to represent, the sequential decision-making problem over multiple time steps in the presence of other interacting agents. The algorithms for solving I-DIDs are haunted by the challenge of an exponentially growing space of candidate models ascribed to other agents over time. In this paper, in order to reduce the candidate model space according the behaviorally equivalent theory, a more efficient way to construct Epsilon behavior equivalence classes is discussed that using belief-behavior graph (BBG). A method of solving I-DIDs approximately is presented, which avoids solving all candidate models by clustering models with beliefs that are spatially close and selecting a representative one from each cluster. The simulation results show the validity of the improved algorithm.

Key Words Agent Modeling, Interactive Dynamic Influence Diagrams (I-DIDs), Dynamic Decision Making, ε-Behavioral Equivalence, Belief-Behavior Graph (BBG)

* 国家自然科学基金资助项目(60975052)
收稿日期: 2011 - 01 - 12; 修回日期: 2011 - 03 - 24
作者简介 李波,女,1981 年生,博士研究生,主要研究方向为多 Agent 系统建模与决策. E-mail: xiaopi_libo@126.com. 罗键,男,1954 年生,教授,博士生导师,主要研究方向为人工智能、多 Agent 系统. 尹华一,男,1980 年生,博士研究生,主要研究方向为物流系统工程、多 Agent 序贯决策. 田乐,女,1981 年生,博士研究生,主要研究方向为 Agent 通信建模.

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1 引　言

多智能体的动态决策是一个相当复杂的决策问题,因其广泛的应用,一直是人工智能研究领域的一个热点. 动态影响图(Dynamic Influence Diagrams, DIDs) [1] 作为单个 Agent 动态决策的建模工具被认为是部分可观察马尔可夫决策过程(Partially Observable Markov Decision Processes, POMDPs) 的一种图形表示方式. 利用 DIDs 描述多 Agent 决策时,要求其它 Agents 行为的概率分布是固定的,这在一定程度上限制了它的应用范围. 合肥工业大学姚宏亮等基于贝叶斯技术和决策理论,在多 Agent 影响图(Multi-Agent Influence Diagrams, MAIDs) 的基础上,提出一种具有更强知识表示能力的动态决策模型: 多 Agent 动态影响图(Multi-Agent Dynamic Influence Diagrams, MADIDs) [2-3] ,用于处理多智能体团队协作问题十分有效. 作为多 Agent 决策问题的建模方法,交互式部分可观察马尔可夫决策过程(Interactive-POMDPs, I-POMDPs) [4] 模型更具有一般性,其适用范围更广. I-POMDPs 通过嵌套结构对环境中的其它 Agent 进行建模,其模型解是在对其它 Agents 行为概率分布的预测下,提供给该 Agent 的最优决策. Doshi, Zeng 等 [5-6] 提出 I-POMDPs 的图形表示形式: 交互式动态影响图(Interactive-DIDs) 模型. I-DIDs 将动态贝叶斯网(Dynamic Bayesian Networks, DBNs) 简洁的表示动态领域知识的能力和影响图(Influence Diagrams, IDs) 具有的决策能力有机地结合起来,为复杂不确定性的动态决策问题,提供一种简洁、直观的知识表示形式,通过引入条件独立性,深入挖掘问题变量之间的隐性结构关系,以此提高问题求解的效率. 该模型突破传统的、基于公共知识的纳什均衡点的假设,具有更为广泛的应用前景.

I-DIDs 的求解算法面临计算上困扰的. 主要原因在于: 对 Agent 建立 I-DID 模型时,其状态空间不仅包含 Agent 所处的物理环境状态,还包括环境中其它 Agents 的候选模型空间,即 Agent 作出行为选择时不仅要考虑自身问题,还要对系统中其它 Agents 进行建模,而其它 Agents 的模型随着时间的增加呈指数级增长,Agent 既要通过推理对环境变化进行预测,同时还要记录其它 Agents 模型随环境变化而演变的全过程,使 I-DIDs 模型求解面临“维度灾难”和“历史灾难”两大挑战. 此外,其它 Agents 模型本身又是 I-DIDs 模型,这种嵌套结构直到 0 层上的 DIDs 模型为止,这更增加了求解的复杂性.

现有的 I-DIDs 求解算法通过限制其它 Agents 候选模型数量,以压缩状态空间,达到简化计算的目的. 文献 [7] ,采用 K-均值方法对候选模型聚类,在每类中选取若干模型作为代表进行更新,最后得到问题的近似解,该方法往往产生一些不必要的模型. Doshi, Zeng 等 [8-9] 提出基于行为等价,动作等价的 I-DIDs 的精确和近似算法,进一步压缩模型空间,取得了不错的效果. 这些算法中,形成行为等价类的过程,需比较候选模型在所有时间片上解的情况,即得到状态空间上所有初始模型所对应的策略树,再自底向上合并策略树, 当其对其它 Agent 的行为预测是完全一致,即策略树一模一样时,方可实现合并. 此过程相当复杂,往往导致计算内存不足,危害 I-DIDs 解法的可升级性.

我们的目的在于找到更有效地获得行为等价类的方法. 本文提出一种基于 ε-行为等价的 I-DIDs 近似求解算法,利用有向无环图表示 Agent 可能的信度—行为,将相同的或信度状态空间接近的模型压缩为同一节点来表示,以实现自顶向下形成行为等价类的方法,这样避免求解各时间片上的所有候选模型,节约了存储空间和计算时间.

2 交互式动态影响图模型及算法

为简单起见,本文主要讨论包含两个 Agent 的交互式动态影响图(I-DIDs) 模型,及其求解算法. 它很容易扩展到包含更多 Agents 的情况.

2.1 I-DIDs 模型

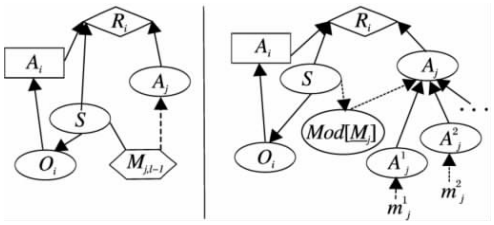
单个时间片上的 I-DIDs 是交互式影响图(I-IDs) [5] . 图 1(a) 是对 i 建立的 l 层上的 I-IDs 模型(l 是 Agents 之间相互建模的嵌套层数) . I-IDs 是由机会节点、决策节点、效用节点和模型节点,以及它们之间的相关连接所组成. 它是在 IDs 的基础上引入了模型节点和策略连接. 其中模型节点(图 1(a) 中的六边形节点 $M_{j,l-1}$) ,用来建模其它 Agent 的所有可能模型; 策略连接(图 1(a) 中带箭头虚线) 如同一个多路连接器,指定了 j 的行为与其模型的对应关系.

模型节点 $M_{j,l-1}$ 是 i 建立的关于 j 的所有可能模型集合. 其中每个模型本身是一个 I-ID 或 ID,这种嵌套结构直到 0 层上的 ID 终止. j 的每个模型可以表示为

$$m_{j,l-1} = \langle b_{j,l-1}, \hat{\theta}_j \rangle,$$

其中 $b_{j,l-1}$ 是表示 j 的信度,$\hat{\theta}_j$ 是 j 的框架,包含动作,观察和效用节点. 为简单起见,假设同一 Agent 不同的两个模型,其信度不同,而模型框架相同. 图 1(b)

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net



(a) 交互式影响模型 (b) 机会节点及相关连接表示的模型节点和策略连接

(a) Model of interactive influence diagram

(b) Representing model node and policy link using chance nodes and dependecies between them

图 1 交互式影响图模型

Fig. 1 Interactive Influence Diagrams

以"平铺"的方式诠释了模型节点和策略连接的含义: 节点 $A_j^q(q=1,2,\cdots)$ 分别对应模型 $m_{j,l-1}^q(q=1,2,\cdots)$ 的最优动作集合,即

$$A_j^q = OPT(m_{j,l-1}^q).$$

此时动作的状态分布可表示为

$$\Pr(a_j \in A_j^q) = \frac{1}{|OPT|}$$

或

$$\Pr(a_j \notin A_j^q) = 0.$$

节点 $Mod[M_j]$ 作为 A_j 的父节点,其状态个数为 $|M_{j,l-1}|$,该节点的概率分布 $\Pr(m_{j,l-1}^q|s)$ 即是给定物理状态 s 的情况下,i 对 j 模型的信度(模型权重),即 $b_i(m_{j,l-1}^q|s)$.

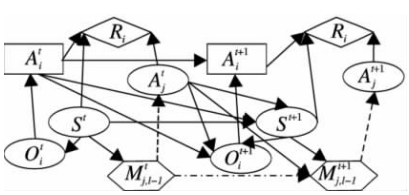


图 2 交互式动态影响图模型

Fig. 2 Interactive Dynamic Influence Diagrams

图 2 给出两个时间片上的 I-DIDs 模型. 结构上,I-DIDs 模型引入模型更新连接(带箭头的点划线). 它包含两阶段的内容: 1) 给定 t 时间片的候选模型集合,确定 $t+1$ 时间片模型节点内存储的模型集合. Agents 执行 t 时刻的动作后,进入 $t+1$ 时间片,并获得新的观察. 考虑到每个模型所对应的最优动作最多有 $|A_j|$ 种,Agent 执行任一动作后获得的观察有 $|\Omega_j|$ 种可能,所以 $t+1$ 时间片的模型个数最多可达 $|M_{j,l-1}^t \| A_j \| \Omega_j|$,其中,$|M_{j,l-1}^t|$ 是 t 时刻模型的个数; $|A_j|$、$|\Omega_j|$ 分别是 Agent j 的动作集合和观察集合中元素个数,节点 $Mod[M_{j,l-1}^{t+1}]$ 的条件概率表(CPT)是关于 $\tau(b_j^t, a_j^t, o_j^{t+1}, b_j^{t+1})$ 的函数,它表示,如果 Agent j 在信度状态为 $b_j^t \in m_{j,l-1}^t$ 时,执行动作 a_j^t,并获得观察 o_j^{t+1},更新信度状态达 $b_j^{t+1} \in m_{j,l-1}^{t+1}$,则函数 $\tau(b_j^t, a_j^t, o_j^{t+1}, b_j^{t+1})$ 的取值为 1,否则为 0; 2) 已知模型的原始信度分布,j 执行的动作和获得的观察,根据标准的贝叶斯推理知识计算 j 模型上新的信度分布,如下公式:

$$\begin{aligned} SE(b_j^t, a_j, o_j) &= b_j^{t+1}(s) \\ &= \beta \sum_{s^t:\, \hat{m}_i^t \hat{\theta}_i^{t+1}} b_j^t(s) \times \\ &\sum_{a_j^t} \Pr(a_i^t|\theta_i^t)\, O_j(o_j^{t+1}, s^{t+1}, a^t) \times \\ &\sum_{o_i^{t+1}} \tau_{\theta_i^{t+1}}(b_i^{t+1}, o_i^{t+1}, a^t, b_i^t) \times O_i(s^{t+1}, \\ &o_i^{t+1}, a^t)\, T_j(s^{t+1}, a^t, st). \end{aligned}$$

图 3 用平铺式的贝叶斯网结构诠释了模型更新的过程. 图中假设

$$|OPT(m_{j,l-1}^{t,1})| = |OPT(m_{j,l-1}^{t,2})| = 1,$$
$$|M_{j,l-1}^t| = 2, \quad |\Omega_j| = 2.$$

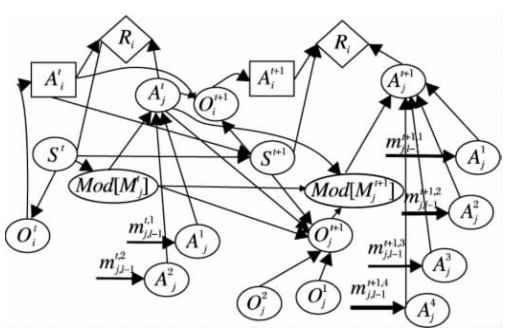


图 3 候选模型更新过程

Fig. 3 Update process of candidate models

若系统中包含更多 Agent,对其中一个建立 I-DID 模型时,只需在图 2 的基础上添加相应的模型节点、策略连接和模型更新连接即可.

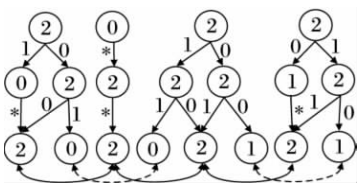
2.2 基于行为等价的精确求解算法

通过上节的介绍,使我们了解到: 求解 I-DIDs 模型关键在于确定模型节点 $M_{j,l-1}^{t+1}$ 中所存储的模型集合. 行为等价原理认为: j 的同一预测行为,对 i 的决策过程影响程度相同,称产生相同预测行为的两个模型为行为等价模型. 因此,i 不必区分处于同一行为等价类中的不同模型,只考虑其中的一个模型不会影响解的最优性. 下面举例说明形成行为等价

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

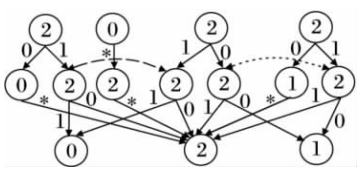
类的过程.

假设某系统中 $|S|=2$, $|A_j|=3$, $|\Omega_j|=2$, $|M^0_{j,l-1}|=4$,求解三个时间片上的 I-DID 模型. 首先获得每一初始模型的解,如图4(a) (* 是通配符,表示任意观察值) ; 然后自底向上合并策略树形成策略图(c) . 行为等价的基本思想要求: 当其它 Agent 的预测行为完全一致时,即策略树一模一样,方可合并对应的两个模型,(c) 图中,虽然前三个模型在 $t=1$ 时刻的最优动作相同,但进入下一时刻后,即使获得相同的观察值却产生不同的动作. 所以它们属于行为不等价的模型,不能进行合并.



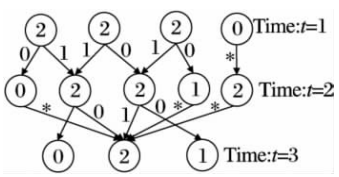
(a) 4 初始模型解的决策树

(a) Policy tree for solving 4 initial models



(b) 合并底层策略树形成的策略图

(b) Policy graph after merging betton of policy trees



(c) 完全策略图

(c) Entire policy graph

图4　获得行为等价类的过程

Fig. 4　Process of obtaining behaviourally equivalent classes

3　基于 ε-行为等价的近似算法

3.1　ε-行为等价和信度 - 行为图

I-DIDs 模型中,Agent 的策略可描述为信度 → 动作的映射,其中信度 $b \in B$ 是环境状态上的概率分布. 事实上,信度状态下值函数有其本身的特性: 有限时间片的 I-DIDs 对应的值函数是一个分段线性凹函数[10]. 这意味着,信度状态空间接近的两个模型,其最优策略可能相同[11]. 根据这一点我们给出 ε-行为等价的定义.

定义1　称集合 $M_{j,l-1}$ 中的模型 $m_{j,l-1}=\langle b_{j,l-1}, \hat{\theta}_j\rangle$ 和 $m'_{j,l-1}=\langle b'_{j,l-1}, \hat{\theta}'_j\rangle$ 为两 ε-行为等价模型,如果它们满足 $\| b'_{j,l-1}, b_{j,l-1} \|_2 \leqslant \varepsilon$,($\varepsilon \geqslant 0$) ,则它们同处在一个 ε-行为等价类中.

由于 i 的决策只受 j 预测行为的影响,所以,没必要区分 j 的产生相同预测行为的2个模型. 文献[5]利用策略图来表示其它 Agent 可能的行为,主要缺点在于计算量过大. 这是因为形成策略图的过程是自底向上合并策略树的过程(如图4) ,需要比较所有候选模型在各时间片上的解的情况. 为更有效地获得其它 Agent 的预测行为,本文提出基于信度 -行为图的表示方法. 下面给出信度 -行为图的定义.

定义2　广义上,信度 -行为图属有向无环图[12] 范畴,但它又具备自身特征,可以定义为一个六元组:

$$BBG=\langle r, V, E, f_v, f_e, SE\rangle,$$

其中 V 是顶点集合,存储 Agent 的信度状态及该信度状态下模型的最优动作集合(B,A) . 设初始顶点为根顶点 $r \in V$,入度为0,存储数据($\mathbb{N}$,B,A) ,$\mathbb{N}$ 是初始时间片上,同一 ε-行为等价类中的模型累积个数,(本文假设初始时间片上候选模型权重相等,即

$$b_i(m^{0,q}_{j,l-1}|s)=\frac{1}{|M^0_{j,l-1}|}, q=1,2,\cdots,|M^0_{j,l-1}|,$$

否则$\mathbb{N}$ 应该表示模型所在 ε-行为等价类中模型权重总和) . $f_v: V\to(B,A)$ 为每个节点分配一个信度状态 B 和一个动作集合 A 中的元素. $f_e: E\to\Omega$: 为每条边分配一个观察集合 Ω 中的元素. f_e 遵循一个重要的特性: 以同一节点为起点的两条边,不能被赋予相同的观察值. 在信度 -行为图中,定义信度转移函数 $SE: B\times A\times\Omega\to B$,$SE(b,a,o)$ 返回 b',$SE(b,v,o)$ 表示在信度为 b 时,执行动作 a,并观察到 o 值时所获得的新的信度.

利用信度 -行为图得到的各时间片上 Agent 的信度集合,即获得各时间片上顶点 V 中的 B 值集合. 定义第 t 个时间步上 Agent 可能的信度集为 B^t,顶点集合 V^t,该顶点集中的元素与根顶点之间的最短通路,表示始点到终点时所观察到的环境信息. 我们用 BBG^T 表示所有时间片上的信度 -行为图. 在信度 -行为图上扩展一个顶点的算法描述如下:

算法1　Extend BBG

输入　B^t_j, V^t_j, a_j

step 1　if($t=0$)

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

step 2 $B_j^0 \leftarrow \Phi, V \leftarrow \Phi$

step 3 for each $b_j^0 \in m_j^0$ in M_j^0

step 4 if(B_j^0 中存在 b_j^0 使 $\| b_j^0 - b_j^0 \| \leqslant \varepsilon$)

step 5 该根顶点 r 中的元素 $\mathcal{N} = \mathcal{N} + 1$

step 6 else

step 7 $B_j^0 \leftarrow b_j^0, a_j \leftarrow f_v(v), V \leftarrow (1, b_j^0, a_j)$

step 8 else

step 9 $B^{t+1} \leftarrow \Phi$

step 10 for each $o_j \in \Omega$

step 11 $b_j^{t+1} \leftarrow SE(b_j^t, a_j, o_j)$

step 12 if(B_j^{t+1} 中存在与 b_j^{t+1} 同处于一个 ε-行为等价类的顶点)

step 13 在父顶点与该顶点间增加一条有向边,权为 o_j

step 14 else

step 15 $B_j^{t+1} \leftarrow b_j^{t+1}, a_j \leftarrow T_p(a_j, o_j), V \leftarrow (b_j^{t+1}, a_j)$

step 16 在父顶点与新顶点间增加一条有向边,权为 o_j

step 17 return B_j^{t+1}, V_j^{t+1}

b_j^{t+1} 表示信度为 b_j^t 时,执行动作 a_j,观察到 o_j,而达到的后续信度,利用标准贝叶斯原理更新信度 $b_j^{t+1} \leftarrow SE(b_j^t, a_j, o_j)$.

3.2 算法描述

本文提出的基于 ε-行为等价的近似算法中,Agent i 根据 j 的信度-行为图判定 j 模型是否需要更新,如算法 1. 这样得到的各时间片上的模型集合是基于 ε-行为等价的最小模型的集合. 由于被处理的模型的顺序不同,得到的最小模型集合不唯一. 另外,算法 1 的 step 12, step 13 指示: 如果下一时间片上该模型的 ε-行为等价模型已经存在,则不需添加新的顶点,即无需对该模型进行更新. 此时,势必会影响 i 对 j 预测行为的概率分布. 为弥补这一损失,需改变 i 对这个已经存在的 ε-行为等价模型的信度,即改变 i 对 j 模型的权重,如下公式:

$$b_i(m_{j,l-1} | s) = b_i(m_{j,l-1} | s) + b'_i(m'_{j,l-1} | s),$$

其中 $b_i(m_{j,l-1} | s)$ 是 i 对已经存在的 ε-行为等价模型的权重,$b'_i(m'_{j,l-1} | s)$ 是 i 对其它与 $m_{j,l-1}$ 同处在一个 ε-行为等价类中模型的权重. 同一 ε-行为等价类中的任一元素都有机会作为该类模型的代表被保留下来,所以,基于 ε-行为等价的最小模型集合不唯一. 基于 ε-行为等价的 I-DIDs 近似求解算法可描述如下:

算法 2 Epsilon BE

初始阶段($t = 0$)

step 1 if $l \geqslant 1$

step 2 for each $m_j^k \in M_{j,l-1}^0$

step 3 反复调用算法 1 Extend BBG

step 4 改变 j 的决策节点 $OPT(m_j^k)$ 成相应的机会节点

step 5 return BBG^T

扩展阶段

step 6 for t from 0 to $T-1$

step 7 if $l \geqslant 1$ then

确定 $M_{j,l-1}^{t+1}$ 中的最小模型集合

step 8 for each $m_j^t \in M_{j,l-1}^t$

step 9 for each $a_j \in OPT(m_j^t)$

step 10 for each $o_j \in \Omega_j$

step 11 更新模型信度 $b_j^{t+1} \leftarrow SE(b_j^t, a_j, o_j)$

step 12 if $b_j^{t+1} \in B_j^{t+1}$

step 13 $M_{j,l-1}^{t+1} \leftarrow m_j^{t+1} = \langle b_j^{t+1}, \hat{\theta}_j \rangle$

step 14 else

step 15 更新节点 $Mod[M_{j,l-1}^{t+1}]$ 的条件概率表,使得行: m_j^t, a_j, o_j,列: 与 m_j^{t+1} 同处在一个 ε-行为等价类的模型 m_j^{t+1},所对应的元素上加 1

step 16 添加模型节点 $M_{j,l-1}^{t+1}$,及节点 $M_{j,l-1}^t$ 与 $M_{j,l-1}^{t+1}$ 之间的模型更新连接

step 17 添加 $t+1$ 时间片上的机会节点,效用节点,决策节点及它们之间的依赖连接,并建立各节点上的条件概率表

结果阶段

step 1 ~ step 17 求解了 $l-1$ 层上的候选模型,即预测了其它 Agent 行为的概率分布,并填充了 l 层 I-DID 所有节点的条件概率表,便可根据标准的动态影响图理论,得到 i 模型的策略树.

3.3 算法复杂性分析

I-DIDs 模型中,t 时间片上,Agent j 的候选模型数量为 $|M_j| = |M_j^0|(|A_j||\Omega_j|)^t$,其中 $|M_j^0|$ 为初始模型数量. 即候选模型空间随时间片的增加呈指数级增长. 基于行为等价原理合并策略树形成策略图,在最坏情况下(不存在可供合并的叶子节点) 的复杂度为 $O((|\Omega|^{T-1})^{|\hat{M}_j|})$,主要由合并策略树时的比较次数决定. 基于 ε-行为等价的近似算法,其最大的优点是实现了自顶向下合并策略树,避免对各时间片上的所有模型进行求解,节省了大量存储空间和计算时间. 该算法中由于同一时间片上不同的两个模型包含的信度之间的距离 $\| b - b' \|_2 \geqslant \varepsilon$,所以各个时间片上形成的模型个数是 $O(\sqrt{n}/\varepsilon)$ 数量级的(n 为信度状态空间维度),有效地控制了候

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

选模型数量,减少了计算量.

4　实验与分析

以 Piotr 等介绍的 Two-Agent 老虎问题为例进行实验. 该问题包含两扇门,左门和右门. 状态集 $S = \{TL, TR\}$,指示老虎在左门后或右门后; 动作集 $A_i = A_j = \{OL, OR, L\}$ 分别表示: 开左门,开右门和倾听. 倾听是收集信息的行为,提供关于老虎位置(老虎的吼叫声) 和其它 Agent 动作(开门声) 的观察信息. 这些动作可以任意组合成团队的联合动作. 联合动作 $\langle L, L\rangle$ 保持环境状态不变,如果任一 Agent 打开一扇门,环境随机地,一致地复原为一个新状态. 观察集 $\Omega 1 = \{GL, GR\}$ 指示老虎的可能位置,其准确率为 0.85,观察集 $\Omega 2 = \{CL, CR, S\}$ 指示其它 Agent 的可能行为其准确率为 0.9,两个观察集中元素可以任意组合成联合观察集合. 当 2 个 Agent 都打开同一扇门且该门后没有老虎时实现最高回报(+20); 当 2 个 Agent 同时打开藏有老虎的门时收到较低的回报(-50); 最差的情况是 Agent 打开相反的门(-100),或者一个 Agent 打开错误的门而另 1 个 Agent 执行 L 动作(-101); 联合动作 $\langle L, L\rangle$ 的代价为 -2. 此外,还要为 0 层上的 DID 模型构造回报函数: Agent 执行 L 的动作获得回报 -1; 打开没有老虎的门获得回报 10,否则为 -100.

4.1　例子

在本例中,取 $\varepsilon = 0.2$,对 i 建立 I-DID 型并求解. 取 $l = 1$, $T = 3$, $|M^0_{j,0}| = 7.0$ 层上,j 的 7 个初始模型,其信度 $\Pr(s = TL)$ 分别为 0.05,0.1,0.15,0.35,0.45,0.85,0.95.

当 $t = 1$ 时,模型 $m^{1,0}_{j,0}$ 对应的信度为 $b^{1,0}_{j,0} = \langle 0.05, 0.95\rangle$,该信度下的最优决策为 OL,建立信度行为图上的第一个顶点 $v^1_1 = (1, \langle 0.05, 0.95\rangle, OL)$. 由于模型 $m^{1,1}_{j,0}$ 与 $m^{1,0}_{j,0}$ 在信度空间上的距离 $\| b^{1,0}_{j,0} - b^{1,1}_{j,0} \| = 0.07 < \varepsilon$,所以不必为该模型建立新的顶点,只需更新 $b^{1,0}_{j,0}$ 所在顶点的数据,此时 $v^1_1 = (2, \langle 0.05, 0.95\rangle, OL)$. 如此反复,获得 $t = 1$ 时的信度行为图(如图 5 顶层上的三个顶点).

当 $t = 1$ 时,对于顶点 v^1_1,根据 Agent 的动作选择及获得的观察,更新模型: $T_p(OL, GL) = T_p(OL, GR) = L$, j 选择"开左门",老虎位置发生重置, $f_e = *$,(其中 * 是通配符,无信息积累), $f_v = (\langle 0.5, 0.5\rangle, L)$,即信度行为图上增加边和顶点 $v^1_2 = (\langle 0.5, 0.5\rangle, L)$,即 $b^{2,0}_{j,0} = \langle 0.5, 0.5\rangle$. 对于顶点 v^2_1, $SE(\langle 0.45, 0.55\rangle, L, GL) = \langle 0.753, 0.247\rangle$,即 $b^{2,1}_{j,0} = \langle 0.753, 0.247\rangle$ 由于 $\| b^{2,0}_{j,0} - b^{2,1}_{j,0} \| = 0.358 > \varepsilon$, $T_p(L, GL) = L$,增加顶点 $v^2_2 = (\langle 0.753, 0.247\rangle, L)$. 同理,依算法 1 获得 j 的信度-行为图,如图 5 所示.

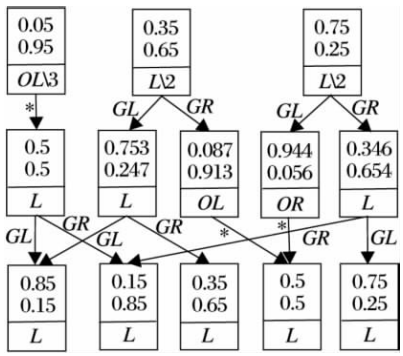


图 5　信度-行为图

Fig. 5　Belief-Behaviour Graph

4.2　结果和分析

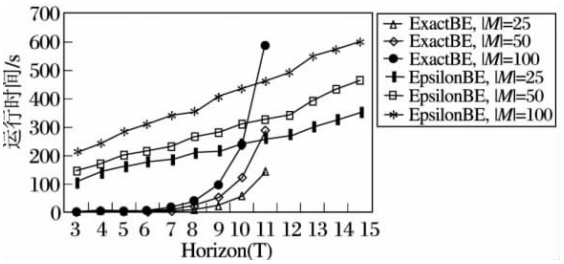
通过比较算法运行时间,Agent 收到的回报,及模型空间大小,来验证本文方法的性能. 在上述 Two-agent 老虎问题和 Two-agent 机器维修问题[10] ($|S| = 3$, $|A_i| = |A_j| = 4$, $|\Omega_i| = |\Omega_j| = 2$),分别利用基于行为等价的精确算法(ExactBE)[9] 和本文中的 EpsilonBE 算法,求解 1 层上的 I-DID 模型. 状态空间中初始模型个数 $|M^0_{j,0}|$ 分别取 25,50 和 100. 平均分配初始模型权重,随机地初始化环境状态. 在 Java 环境下运行程序 50 次,取平均值作为最后结果. 系统配置: WinXP, Dual processor 1.73GHz, 2GB memory.

图 6 用双轴曲线绘制了两种算法形成行为等价类的运行时间((a),(b) 分别为老虎和机器维修问题上的实验结果). 实验数据表明,ExactBE 算法中形成行为等价类的运行时间随时间(度量单位: s) 增长迅速,而 EpsilonBE(本例取 $\varepsilon = 0.01$) 算法的运行时间(度量单位: millisec/10) 基本呈线性增长规律. 图 7((a),(b)) 给出 ε 上述两个问题 10 个时间片,不同 ε 值下运行 EpsilonBE 和 ExactBE 算法后,所获得的平均收益的近似和精确值.

图 7 显示, ε 值越小,Agent 所获得的回报越接近问题的精确解. 直觉上,算法所需的运行时间就越长.

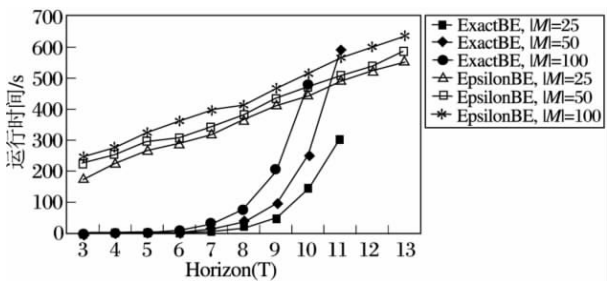
表 1 给出了取不同 ε 值,运行两种算法,各时间片上的模型节点所保留的模型个数. ExactBE 算法,候选模型个数随时间片的增长而增加; 而 EpsilonBE 算法,所保留的模型个数主要受 ε 值影响,上下波动

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net



(a) 老虎问题的运行时间曲线

(a) Running time curves on tiger problem

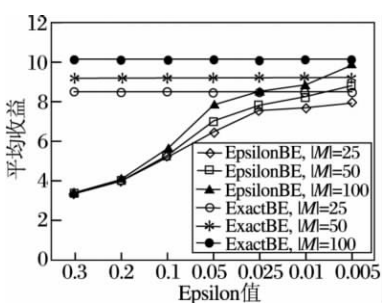


(b) 机器维修问题的运动时间曲线

(b) Running time curves on machine maintenance problem

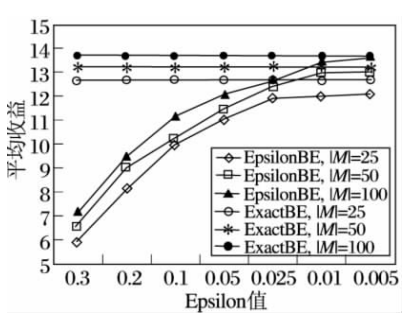
图 6 ExactBE 与 EpsilonBE 运行时间曲线

Fig. 6 Curves of running time for ExactBE and EpsionBE



(a) 老虎问题平均收益曲线

(a) Average reward curve on tiger problem



(b) 机器维修问题平均收益曲线

(b) Average reward curve on machine maintenance problem

图 7 平均收益曲线

Fig. 7 Average reward curves

表 1 各时间片上的模型节点所保留的模型个数

Table 1 Number of models maintained in the model node for different time horizon

Level 1	*T*	模型节点中的模型个数			
		ExactBE	EpsilonBE		
			0.2	0.1	0.05
Two-Agent 老虎问题	4	5	4	8	10
	7	8	3	5	6
	10	11	5	7	8
	13	12	5	7	5
	15	14	4	5	9
Two-Agent 机器维修问题	4	5	5	7	10
	10	12	4	6	8
	15	17	5	8	9

不大. 正是如此,I-DIDs 模型的时间片 越大, EpsilonBE 算法所表现的优越性越明显.

5 结 束 语

本文基于 ε-行为等价的 I-DIDs 改进算法,利用信度-行为图,通过比较候选模型中的信度距离指导模型是否需要更新,实现自顶向下形成各时间片上的近似行为等价类,避免对候选模型空间上的所有模型进行求解,节省了存储空间和计算时间,提高了 I-DIDs 模型求解问题的能力. I-DIDs 模型的解是在预测其它 Agents 行为概率分布的基础上提供给该 Agent 的最优决策,能更有效地描述多 Agent 决策问题. 但是,该模型忽略了 Agents 之间的通信能力,在一定程度上限制了该模型的应用范围,也增加了求解过程的复杂性. 因此,进一步的研究工作要考虑包含通信行为的 I-DIDs 模型(Com-I-DIDs) ,及其求解算法.

参 考 文 献

[1] Tatman J A, Shachter R D. Dynamic Programming and Influence Diagrams. IEEE Trans on Systems, Man and Cybernetics, 1990, 20: 365 - 379

[2] Yao Hongliang, Wang Hao, Zhang Yousheng, *et al*. Multi-Agent Dynamic Influence Diagrams and Its Approximation of Probability Distribution. Pattern Recognition and Artificial Intelligence, 2007, 20(4): 521 - 532 (in Chinese)
(姚宏亮,王 浩,张佑生,等. 多 Agent 动态影响图及其概率分布的近似方法. 模式识别与人工智能, 2007, 20(4): 521 - 532)

[3] Yao Hongliang, Wang Hao, Wang Ronggui, *et al*. Approximate Computation of Multi-Agent Dynamic Influence Diagrams. Journal of Computer Research and Development, 2008, 45(3): 487 - 495 (in Chinese)
(姚宏亮,王 浩,汪荣贵,等. 多 Agent 动态影响图的近似计算方

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

法. 计算机研究与发展, 2008, 45(3): 487 – 495)

[4] Gmytrasiewicz P J, Doshi P. A Framework for Sequential Planning in Multi-Agent Settings. Journal of Artificial Intelligence Research, 2005, 24(1): 49 – 79

[5] Doshi P, Zeng Y F, Chen Q Y. Graphical Models for Interactive POMDPs: Representation and Solutions. Journal of Autonomous Agents and Multi-Agent Systems, 2009, 18(3): 376 – 416

[6] Polich K, Gmytrasiewicz P J. Interactive Dynamic Influence Diagrams // Proc of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems, New York, USA: ACM Press, 2007: 147 – 149

[7] Zeng Y F, Doshi P, Chen Q Y. Approximate Solutions of Interactive Dynamic Influence Diagrams Using Model Clustering // Proc of the 22nd International Conference on Association for the Advancement of Artificial Intelligence. Vancouver, Canada: AAAI Press, 2007: 782 – 787

[8] Zeng Y F, Doshi P. Speeding up Exact Solutions of Interactive Dynamic Influence Diagrams Using Action Equivalence // Proc of the 21st International Joint Conference on Artificial Intelligence. Pasadena, USA, 2009: 1996 – 2001

[9] Doshi P, Zeng Y F. Improved Approximation of Interactive Dynamic Influence Diagrams Using Discriminative Model Updates // Proc of the 8th International Conference on Autonomous Agents and Multi-Agent Systems. Budapest, Hungray, 2009: 907 – 914

[10] Smallwood R D, Sondik E J. The Optimal Control of Partially Observable Markov Decision Processes over a Finite Horizon. Operations Research, 1973, 21(5): 1071 – 1088

[11] Pynadath D V, Marsella S C. Minimal Mental Models // Proc of the 22nd International Conference on Association for the Advancement of Artificial Intelligence. Vancouver, Canada, 2007: 1038 – 1044

[12] Geng S Y, Qun W L. Discrete Mathematics. Beijing: Higher Education Press, 1998

(耿素云,屈婉玲. 离散数学. 北京: 高等教育出版社,1998)

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net