动态影响图模型研究

俞 奎[1,2],王 浩[2],姚宏亮[2]
(1. 常州纺织服装职业技术学院,江苏 常州 213164;2. 合肥工业大学 计算机与信息学院,安徽 合肥 230009)

摘 要:部分可观察马尔可夫决策过程在策略空间和状态空间上的计算复杂性,使求解其一个最优策略成为 NP-hard 难题. 为此,提出一种动态影响图模型来建模不确定环境下的 Agent 动态决策问题. 动态影响图模型以有向无环图表示系统变量之间的复杂关系. 首先,动态影响图利用动态贝叶斯网络表示转移模型和观察模型以简化系统的状态空间;其次,效用函数以效用结点的形式清晰地表示出来,从而简化系统效用函数的表示;最后,通过决策结点表示系统的行为来简化系统的策略空间. 通过实例从 3 个方面和 POMDP 模型进行了比较,研究的结果表明,动态影响图模型为大型的 POMDP 问题提供了一种简明的表示方式,最后在 Robocup 环境初步验证了该模型.

关键词:动态贝叶斯网络; 影响图; 马尔可夫决策过程; 部分可观察马尔可夫决策过程; 动态影响图
中图分类号: TP181 **文献标识码**: A **文章编号**: 1673-4785(2008)02-0159-08

A dynamic influence diagram for dynamic decision processes

YU Kui[1,2], WANG Hao[2], YAO Hong-liang[2]
(1. Department of Computer Science, Institute of Textile and Garment of Changzhou, Changzhou 213164, China; 2. School of Computer and Information, Hefei University of Technology, Hefei 230009, China)

Abstract: Computational complexities in strategy space and state space make the partially observable Markov decision process (POMDP) an NP-hard problem. Therefore, in this paper, a dynamic influence diagram is proposed to model the decision-making problem with a single agent, in which a directed acyclic diagram is used to express the complex relationships between systematic variables. Firstly, a dynamic Bayesian network is used to represent the transition and observation models so as to reduce the state space of the system. Secondly, in order to reduce the representational complexity of the utility function, it is expressed in terms of utility nodes. Finally, the actions of the system are represented with decision nodes to simplify the strategy space. The dynamic influence diagram is compared with the POMDP using these three aspects. Our research indicates that a dynamic influence diagram provides a simple way to express POMDP problems. Experiments in the Robocup environment verified the effectiveness of the proposed model.

Keywords: dynamic Bayesian networks; influence diagrams; Markov decision process; partially observable Markov decision process; dynamic influence diagram

动态决策问题是人工智能领域研究的中心问题之一. 马尔可夫决策过程[1] (Markov decision process, MDP)和部分可观察马尔可夫决策过程[2] (partially observable Markov decision process, POMDP)是研究动态决策问题的有效模型. MDP 所假设的是 Agent 能够确切获知环境的状态. POMDP 通过从环境中获得的观察来感知环境的状态. 因此 POMDP 更适合不确定情形下动态决策问题的建模. 求解 POMDP 的算法通常是把 POMDP 转换为一个连续状态下的 MDP 求解. Kaelbling 等人给出了求解 POMDP 的精确求解算法[3],由于信度状态 MDP(belief MDP)是一个连续状态的模型,策略空间和状态空间的复杂性呈指数级增长,算法实际上不可行. 因此许多研究者提出了 POMDP 的近似求解算法[4-5]. 实际上对于离散的 POMDP,求

收稿日期:2007-06-20.
基金项目:国家自然科学基金资助项目(60575023,60705015);安徽省自然科学基金资助项目(070412064).
通讯作者:俞 奎. E-mail:ykui713@hotmail.com.

解一个最优策略是 PSPACE-complete 难题[6];求解 εoptimal 策略是 NP-hard 难题[7]. 状态空间、值函数和策略表示的复杂性是造成求解 POMDP 问题的主要难点.

本文在动态贝叶斯网络和影响图的基础上提出一种动态影响图(dynamic influence diagram, DID)模型来建模不确定环境下的 Agent 动态决策问题,以解决 POMDP 模型不能有效求解复杂的动态决策问题. DID 是一个有向无环图,用 DBN 表示转移模型和观察模型,用效用节点表示效用函数,决策节点表示系统的行为,这样 DID 利用图模型蕴涵的条件独立性简洁表示了系统中变量之间的复杂关系.

动态贝叶斯网络[8] (dynamic bayesian networks, DBN)是贝叶斯网络的一种扩展模型,是对具有随机过程性质的不确定性问题进行建模的一种有力的工具. 由于 DBN 没有决策节点,DBN 不易处理 Agent 的决策问题.

影响图[9] (influence diagrams, ID)是贝叶斯网络的另一种扩展模型,影响图可以同贝叶斯网络一样,表示多变量之间的复杂关系;另外,由于其结合了效用理论,因而比贝叶斯网络更适合处理 Agent 决策. 它虽然在贝叶斯网络基础上结合了效用理论,但不易表示和处理动态的决策问题.

DID 将 DBN 简洁的表示动态领域知识的能力和影响图具有的决策能力有机地结合起来,使 DID 从状态空间、值函数的表示和策略空间 3 个方面降低了求解动态决策问题的复杂性.

1) DBN 表示 DID 的转移模型和观察模型. DID 利用 DBN 蕴涵的条件独立性和稀疏性简化系统的状态空间;

2) 效用函数用效用结点的形式清晰地表示出来,简化系统的效用函数的表示;

3) DID 中的决策结点表示系统的行为,一个决策节点只和少数几个状态变量相联系,简化系统的策略空间.

研究表明 DID 为大型的 POMDP 问题提供了一种简明的表示方式.

1 POMDP 模型

MDP 是表示和处理单个 Agent 动态决策问题的主要模型. MDP 由四元组 $< S, A, R, T >$ 定义,S 是一个环境状态集,A 表示系统行为集,奖赏函数 $R: S \times A \to R$ 和状态转移函数 $T: S \times A \to \mathrm{PD}(S)$. 记 $R(s, a, s')$ 为系统在状态 s 采用 a 动作使环境状态转移到 s' 获得的瞬时奖赏值;记 $T(s, a, s')$ 为系统在状态 s 采用 a 动作使环境状态转移到 s' 的概率. MDP 模型所假设的是 Agent 能够确切获知环境的状态.

POMDP 不要求 Agent 事先了解身处环境的状态,而是通过从环境中获得的观察来感知环境的状态. POMDP 模型的定义如下[2]:

POMDP 由六元组 $< S, A, T, R, Z, O >$ 定义. 其中:1) S:环境状态的有限集合;2) A:动作的有限集合;3) $T: S \times A \to \Pi(S)$;其中 $\Pi(S)$ 为状态的分布,在某一状态 s 执行动作 a 后,下一个状态 s' 的概率分布,一般用 $T(s, a, s')$ 表示;4) $R: S * A \to R$;报酬函数,表示在状态 s 执行动作 a 所能获得的报酬,也用 $R(s, a)$ 表示;5) Z:观察的有限集合,即系统可以感知的世界状态集合;6) $O: S * A \to \Pi(Z)$;其中 $\Pi(Z)$ 为观察的分布. 观察函数,表示系统在采取动作 a 转移到状态 s' 时,观察函数 O 确定其在可能观察上的概率分布,记为 $O(s', a, o)$.

$$p(b, a) = \sum_{s \in S} R(s, a)\, b(s).$$

由于 POMDP 问题最优策略学习转变为"信度状态 MDP"(belief MDP)最优策略的学习,造成一个严重的问题就是由于其状态空间是连续的,值函数、策略表示和状态空间的复杂性使求解 POMDP 是一个 NP-hard 难题.

因此针对状态空间和值函数表示的复杂性,Boutilier 和 Poole 把因式 MDP 扩展应用到 POMDP 模型中,提出了因式 POMDP (factored POMDP)[10]. 因式 POMDP 利用动态贝叶斯网络来表示转移模型和观察模型,用决策树或代数决策图来表示转移概率和值函数. 由于 POMDP 模型的值函数是分段线性凸函数,随着状态空间的增长,表示值函数的决策树或代数决策树的数目会指数级的增长. 因此,因式 POMDP 仍然不能有效处理大型的 POMDP 的最优决策问题.

在 Boutilier 等人研究的基础上,把动态贝叶斯网络和影响图结合,提出动态影响图(dynamic influence diagram, DID)模型来建模不确定环境下的 Agent 动态决策问题,以解决复杂情况下的动态决策问题.

2 动态影响图模型

2.1 动态影响图模型的表示

动态影响图可以定义成一个二元组 $< B_t, B_{ts} >$,其中 B_t 表示 t 时刻的影响图,$B_{ts} = (E_t, V_{ts})$,其中 E_t 表示连接 B_t 和 B_{t-1} 的有向边,V_{ts} 表示 E_t 所连接的顶点的集合. 动态影响图结点集合 $V =$

(D, X, O, U),其中 D 表示决策变量集,X 表示随机变量集,O 表示 X 所对应的观察值变量集,U 表示效用结点集. 设在 t 时刻系统的状态随机变量集为 $X = \{X_1, X_2, \cdots, X_n\}$,状态对应的观察随机变量集为 $O = (O_1, O_2, \cdots, O_n)$,决策结点变量集为 $D = \{D_1, D_2, \cdots, D_n\}$,每个决策结点有一个与行为集对应的效用结点集为 $U = \{U_1, U_2, \cdots, U_m\}$. 一个 DID 模型有结构策略模型、概率模型和效用模型 3 个部分构成.

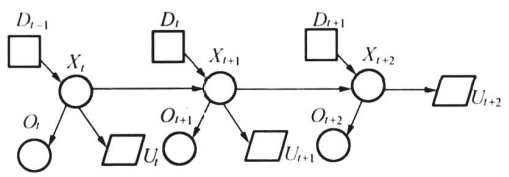


图 1　单个 Agent 3 个时间片的动态影响图

Fig. 1　Three time slices DID for a single Agent

定义 1　结构策略模型　设 $\Lambda_t = (\delta_1, \cdots, \delta_t^n)$ 是 t 时刻 Agent 的一个决策规则集,结构策略模型为 t 时刻的每个决策结点 D_t 确定一个局部决策规则 δ_t,δ_t 是 D_t 的父结点集 $Pa(D_t)$ 和行为选择之间的一个映射:

$$\delta_t : \Omega_{Pa(D_t)} \rightarrow \Omega_{D_t}. \tag{1}$$

式中:$\Omega_{Pa(D_t)}$ 表示决策结点 D_t 的父结点集可能的取值空间,Ω_{D_t} 表示决策结点 D_t 可能的行为空间. 一个决策规范 σ_t 是为 t 时刻的决策集中每个决策结点分配一个决策规则.

定义 2　概率模型　给定决策规则时决策结点看成随机结点,给定 t 时刻的决策规则 δ_t,决策结点 D_t 的条件概率为

$$P_{\delta_t}(D_t \mid Pa(D_t)) = \begin{cases} 1, \delta_t(Pa(D_t)) = d_j; \\ 0, \text{其他}. \end{cases} \tag{2}$$

式中:d_j 表示为 Agent 行为集合中的一个行为.

在动态影响图中,转移模型和观察模型可以表示为一个 DBN,系统状态被一组随机变量 $X = \{X_1, X_2, \cdots, X_n\}$ 表示,每个 X_i 在有限值域 $\mathrm{Dom}(X_i)$ 中取一定的值. 每个变量 X_i 的取值定义了系统的一个状态. 令 X_i 是当前时间片的变量,X_i' 是下一个时间片的变量. X_i' 的父结点集为 $\mathrm{Parents}(X_i') \subseteq (X, D)$. 每个结点 X_i' 有一个条件概率表 $Pa(X_i' \mid \mathrm{Parents}(X_i'))$. 给定决策规则 δ_t,转移概率 $P_{\delta_t}(x' \mid x, d)$ 表示为

$$P_{\delta_t}(x' \mid x, d) = \prod P_{\delta_t}(x_i' \mid u_i).$$

式中:u_i 是变量 x 和 d 在 $\mathrm{Parent}_{\delta_t}(X')$ 中的取值.

可观察变量集 $O = \{O_1, O_2, \cdots, O_n\}$,在动态影响图中假定 $\mathrm{Parents}(O_i) \subseteq X'$,即执行每个行动后可观察变量的取值依赖于系统状态变量,同时假定可观察变量没有孩子结点. 则给定决策规则 δ_t 动态影响图的观察模型为 $P_{\delta_t}(O_i \mid u_i')$,$u_i'$ 是变量 x 在 $\mathrm{Parents}(O_i')$ 中的取值.

令 $e[t]$ 是动态影响图在 t 时刻可观察变量 O 的一组取值,$e_{1:t}$ 为到 t 时刻可观察变量的所有可能取值,$x[t]$ 为 $X[t]$ 的取值,在给定决策规则 δ_t 时,则随机变量、观察值变量和决策变量在 t 时刻的联合概率分布表示为

$$P_{\sigma_t}(X_t, E_t, D_t) = \prod_{X_t \in X_t} P(X_t \mid X_{t-1}, D_{t-1}) \prod_{X_t \in X_t} P(O_t \mid X_t) \prod_{D_t \in D_t} P_{\delta_t}(D_{t-1} \mid pa(D_{t-1})).$$

式中:X_t 是 Agent 的状态变量集,O_t 是观察值的变量集.

定义 3　效用模型　在动态影响图中用一个效用结点 U_t 表示 t 时刻状态的效用值. 每个效用结点 U_t,有一个和父结点集 $P_a(U_t)$ 相联系的局部效用函数 $U_t(P_a(U_t))$. 由于效用函数具有时分性,在 DID 里每个时间片包含一个回报结点,则各个局部效用和为

$$U(X, D) = \sum_{t=1}^{m} U_t(P_a(U_t)). \tag{3}$$

每个效用结点和一个局部效用函数相联系,一个效用结点 U_t 的效用函数可表示为

$$U_t(P_a(U_t)) = f_t(X_t^1, \cdots, X_t^n) = w_t^1 X_t^1 + \cdots + w_t^n X_t^n.$$

式中:$P_a(U_t) = \{X_t^1, \cdots, X_t^n\}$ 是效用结点 U_t 父结点集;权重 w_t^i 对应一个变量 $X_t \in P_a(U_t)$,其表示 X_t 对效用结点 U_t 影响的度量.

则对于给定策略规范 σ 时,时刻 t 的期望效用为

$$\mathrm{EU}(\sigma_t) = \sum_{X_t} P_{\sigma_t}(X_t, O_t, D_t) \sum_{U_t \in O_t} (P_a(U_t)). \tag{4}$$

2.2　动态影响图模型的性质

定理 1　动态影响图中,给定决策规则 δ_t 时,$P_{\delta_t}(x[t+1] \mid e[t+1], d_t)$ 的信度更新与 POMDP 中相同决策规则下的 $b'(s')$ 是等价的.

证明　令在 t 时刻 DID 执行给定决策规则 δ_t,得到观察 $e[t+1]$ 时,$P_{\delta_t}(x[t+1] \mid e[t+1], d_t)$ 的信度更新如下:

1) 预测

$$P_{\delta_t}(x[t+1] \mid e_{12}, d_{12}) = \sum_{x[t]} P(x[t+1] \mid x[t] d_2, e_{12}) P(x[t] \mid e_{12}, d_{12}) = \sum_{x[t]} P(x[t+1] \mid x[t], d_2) P(x[t] \mid e_1, d_{t-1}).$$

(5)

2) 更新

$$P\delta_t(x[t+1] \mid e[t+1], d_t) =$$

$$P(x[t+1] \mid e_{12}, e[t+1], d_t) =$$

$$\partial P(e[t+1] \mid x[t+1])\, e_{12})\, P(x[t+1], e_{12}, d_2) =$$

$$\partial P(e[t+1] \mid x[t+1])\, P(x[t+1], e_{12}, d_2) =$$

$$\partial P(e[t+1] \mid x[t+1])$$

$$\sum_{x[t]} P(x[t+1] \mid x[t] \mid e_{12}, d_{12}. \qquad (6)$$

式中: ∂是归一化因子, $P(e[t+1] \mid x[t+1])$ 为观察模型,式(6)中的求和式中,第1个因子是转移模型,第2因子则是当前状态分布. 由式(5)和(6)可以用任何的DBN推理算法求解计算 $P(x[t+1] \mid e[t+1], d_t)$,这样就可以利用DBN强大的概率推理能力进行信度更新计算.

在POMDP模型中,假定在当前状态 s 下执行行动 a,得到观察 o,则到达状态 s' 的信度 $b(s')$ 为

$$b(s') = P_r(s' \mid o, a, b) =$$

$$\partial P_r(o \mid s', a, b)\, P_r(s' \mid a, b) =$$

$$\partial P_r(o \mid s', a) \sum_s P_r(s' \mid a, b, s)\, P_r(s \mid a, b) =$$

$$\partial O(s', a, o) \sum_s T(s, a, s')\, b(s). \qquad (7)$$

式中: $O(s', a, o)$ 是观察模型, $T(s, a, s')$ 转移模型; $b(s)$ 是系统的当前状态. 显然与式(6)是等价的. 但将复杂系统的状态分解成一些组成变量,利用DBN中条件独立性和稀疏性简化系统的状态空间数. 因此式(6)和(7)不仅是等价的,而且式(6)利用条件独立性指数级地降低了式(7)的参数个数和状态空间数,证毕.

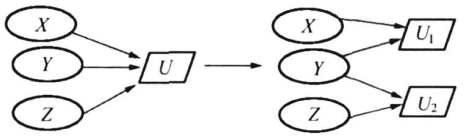
定义4 效用函数的时分性(time-separability)是指一个状态序列的效用是状态序列里的每个状态的效用的和. 累加可分性(additive separability)指一个效用函数可分解为一组效用函数之和.

定理2 动态影响图中的效用结点具有时分性和累加可分性.

证明 在DID中,效用函数是由效用结点清晰的表达,每个时间步中的效用函数是条件独立的,如图1,在 $t+2$ 时间步的效用是前面每个时间步获得的期望回报累加和. 这与POMDP中的回报函数具有时分性是一致的.

在文献[11]中,Tatman和Shachter证明了在影响图的多步决策问题中效用结点的累加可分离性,如图2.

由于DID每个时间片其实就是一个影响图,如图1. DID可以近似为一个影响图的多步决策问题的扩展,因此DID中的效用结点也具有累加可分性.



$$U(X, Y, Z) = U_1(X, Y) + U_1(Y, Z)$$

图2 一个影响图效用结点的分解

Fig. 2 The separability of utility node

定理3 给定一个DID的初始决策规范 σ_0 和初始效用函数 $U_o(X_0, D_0)$,则 $t=1$ 时的 σ_1 和 $U_1(X_1, D_1)$ 在时间上具有不变性.

证明 由式(1)和式(6)可知, $t=1$ 时的 σ_1 和 $U_1(X_1', D_1)$ 完全依赖于 B_t 和 B_{ts} 的结构,又因为DID的 B_t 和 B_{ts} 结构具有时间不变性,因而 $t=1$ 时的 σ_1 和 $U_1(X_1, D_1)$ 在时间上也具有不变性. 而 $t=0$ 时的 σ_0 和 $U_o(X_0, D_0)$ 是和时间无关的,需要事先给定,证毕.

3 模型比较

DID以有向无环图简洁地表示系统中变量之间复杂的关系. 本节从知识表示与状态空间、模型的值函数的表示、模型推理能力3个方面把DID模型和POMDP模型进行比较.

3.1 知识表示与状态空间

POMDP模型通过列举系统可能所有的状态为其建模,存储每个状态的转移概率和其回报函数值. DID利用图模型中蕴涵的条件独立性来分解联合概率分布,降低知识表示和获取的复杂性. DID将复杂系统的状态分解成一些组成变量,以简洁的图模型表示了系统中主要的变量,清晰地描述了系统变量之间的关系. 利用DBN蕴涵的条件独立性和稀疏性,转移模型和观察模型的全联合分布可分解成若干个局部的概率分布表示,同时网络的拓扑结构表明如何从局部的概率分布获得全联合分布,大大简化了系统的状态空间.

例如一个感冒病人的症状可以用2个变量来描述: headaches 和 temperature ,令 $X = \{H(\text{headaches}), T(\text{temperature})\}$, $H = (\text{yes}, \text{no})$, $T = (\text{normal}, \text{slightly high}, \text{high})$. 则2个时间片中变量 H 和 T 是相互独立的,即 $P(T_i \mid T_{i-1}, H_{i-1}) = P(T_i \mid T_{i-1})$. 这样在DID只需要 $4+9=13$ 个转移概率参数,而POMDP需要 $5 \times 5 = 25$ 个转移概率参数,假设行为的状态空间为1.

一般情况下,假设状态空间 Ω_x 可用 M 个状态变量表示,每个变量至多有 S 个状态和 N 个父结点. 则利用条件独立性联合概率分布可用 $O(M(S^N))$ 参数,否则需要 $O(S^M)$ 个参数. 假设 $M=10$, $S=3$, $N=4$,需要 810 个参数而不是 5 904 900 个参数. 下面考虑一个复杂一点的例子,如图 3.

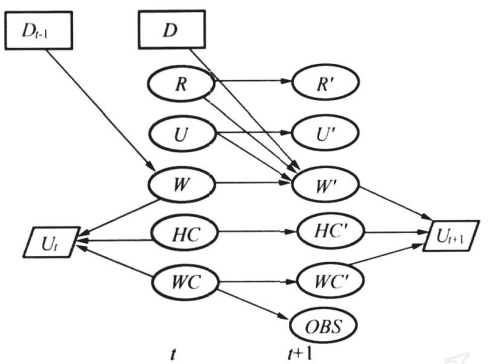


图 3　一个动态影响图

Fig. 3　A example of DID

图 3 是一个动态影响图的例子. 如果顾客想要咖啡(want coffee, WC),并得到了咖啡(has coffee, HC)机器人得到奖赏(reward, R),否则得到惩罚. 机器人如果在下雨(rain, R)的情况下穿过街道去买咖啡(get coffee, GetC)的话,可能被淋湿(wet, W),除非带伞(umbrella, U),当然机器人可以完成其他的任务,详见文献[2]. 这里是一个简化的向前看一步的动态影响图,图中只标出了时间片之间变量的关系. OBs 是一组可观察变量, D_t 是决策(行为)结点. 该例子有 5 个布尔型的状态变量: R、U、W、HC、WC 和一个决策变量 D, U 为效用结点,假设 $|\mathrm{OBs}|=2$, $|A|=2$.

在图 3 中, $|S|=2^5=32$, $|O|=2$, $|A|=2$,则 POMDP 模型需要的转移概率参数为 $|S|\times|S|\times|A|=2\ 048$,观察模型需要的概率参数为 $|O|\times|S|\times|A|=128$,这样使 POMDP 只适合处理小规模的问题. 在 DID 模型中 $P(R',U',W',HC',WC'|R,U,W,HC,WC,D_t)=P(R'|R)P(U'|U)P(W'|R,U,W,D_t)P(HC'|HC)P(WC|WC')$. 这样就把整个系统变量的全联合分布转变成局部的概率分布表示,图 3 中所需的转移概率参数转换为 5 个局部概率表示,只需要 40 个参数,因此 DID 大大简化了系统的状态空间.

3.2　模型的值函数表示

由于求解 POMDP 问题的算法通常把 POMDP 转换为一个连续状态下的 MDP 求解,即信度状态 MDP,因此其状态空间是连续的, POMDP 的最优值函数用分段线性凸函数表示,线性分段凸函数由一组 a 向量表示,如图 4.

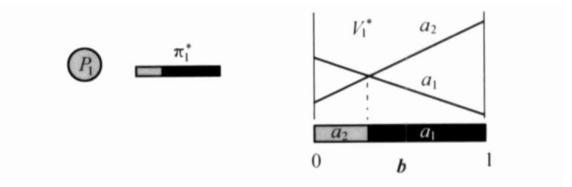


图 4　2 个状态的 POMDP 模型的值函数 V_1^*

Fig. 4　Value function V_1^* of POMDP for two states

根据定义,信状态度 MDP 模型是一个具有连续状态空间的模型,精确求解算法的复杂度非常高,以至于求解状态、动作和观察的集合元素超过 10 的问题都非常困难[6]. 一个重要的原因就是用来表示最优值函数的线性分段凸函数,这种表示形式的复杂度随着迭代的进行,增长得也非常得快. 表示第 $i+1$ 步值函数的 a 向量集的大小会在一步迭代后就达到指数量级 $O(|A||\Gamma_i|^{|\Omega|})$,这一步更新所需的计算复杂度是 $O((|A||S|^2|\Gamma_i|^{|\Omega|})$,其中 $|\Gamma_i|$ 表示第 i 步值向量的数目, $|\Omega|$ 表示状态空间. 如果假设初始的值函数是线性的,那么经过 i 次迭代后,第 i 次值函数对应的向量集大小是 $O(|A|^{|\Omega|^{i-1}})$,即使对于简单的有限阶 POMDP 问题,求解最优策略的时间复杂度也是 PSPACE-hard 问题. 无限阶的问题求解则可能是不可计算的. 一种解决办法是求解与最优解差距小于某个特定精度的近似解,但有时即使是这样都是不可计算的[7]. 因此,大量的研究重点被放在寻找近似解的算法研究上,实际上求解 εoptimal 策略也是 NP-hard 难题.

在 DID 中效用函数以效用结点的形式清晰地表达出来,对于离散的 DID 模型,其变量的状态空间仍然是离散的,避免了用分段线性凸函数表示效用函数. DID 中的效用节点并不依赖所有的变量,通常只由一部分变量决定. 如图 3,影响效用结点是 W、WC、HC,而与其他节点无关. 这样大大简化了效用函数表示的复杂性,但其复杂性会随着其父节点的个数的增加而指数级的增加. 但幸运的是由于父节点的个数总是有限的,又由定理 2 可知,效用函数的累加可分性可进一步分离效用函数,使每个效用结点依赖于更小的一组状态和行为变量,如图 5.

在 POMDP 中,值函数被转化为与每个状态转移相联系的效用值. 更改值函数可能需要对所有的转移和状态的回报函数重新构造. 在 DID 中效用函

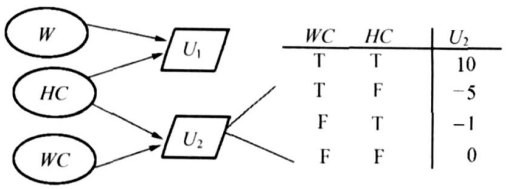


图5 累加回报:$R(W, HC, WC) = U_1(W, HC) + U_2(HC, WC)$

Fig. 5 Addictive rewards: $R(W, HC, WC) = U_1(W, HC) + U_2(HC, WC)$

数通过特定的结点明确地表示出来,并且效用结点只依赖对其影响的状态变量. 这样可以根据特定的问题更改效用函数,而在 *POMDP* 模型中则是不行的. 这个特点不仅可以容易进行系统的敏感性分析,而且可以根据环境的需要方便地更改效用函数,进行验证和改进模型.

一个大型的 POMDP 模型可能有几百万个状态和几十个行为,价值迭代或策略迭代算法在计算各个状态效用时要考虑每一个状态和行为,实际上并不是每个行为都影响系统所有的状态. 因此,DID 把系统的行为分解成一组决策变量,用决策节点表示系统的每个行为. 每个决策节点都有几个状态,每个状态表示系统的一个行动. 利用条件独立性,这样决策节点只与部分状态变量相联系. 如图 3,决策节点只影响状态变量 W,在计算 $t+1$ 时刻 D_t 最优决策时,只需要根据 W、HC、W 等 3 个状态变量计算不同的策略下的效用值,这样可大大简化策略空间的搜索.

3.3 模型的推理

每个时刻的信度更新计算是求解 POMDP 模型最优策略的关键环节,但由于 POMDP 状态空间较大,信度更新计算需要处理庞大的转移概率矩阵. 由式(7)可以得出一个状态的信度更新时间为 $O(|S|^2|O|)$ 这使得 POMDP 模型中的推理代价太大.

而在 DID 中,DBN 表示的转移模型和观察模型的概率分布以局部的概率分布表示出来,已大大简化了系统的状态空间. DBN 作为一种时序概率模型,已有许多较成功的概率推理算法,如联合树算法、BK 和粒子滤波算法等. 在给定策略情况下,决策节点的状态已知,除去效用节点(效用节点对信度更新没有影响),DID 实质是一个 DBN,因此,这样可以利用 DBN 强大的概率推理能力进行 DID 的信度更新计算. 因此在信度更新的复杂性方面,DID 和 POMDP 模型之间的关系可以类比于 DBN 和隐马尔可夫模型之间的关系. 在一个给定策略下,图 6 是把图 1 的 DID 去掉效用节点后转换成一个 DBN.

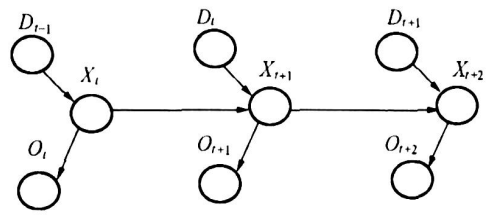


图6 3个时间片的 DBN

Fig. 6 A DBN for three slices

另外,从模型的可学习能力来说,DID 的转移模型和观察模型,可利用 DBN 的参数和结构学习算法来学习它们的参数和结构. 而 POMDP 模型中的转移概率是全联合分布,状态之间的关系是全连接关系,学习如此巨大的参数是非常困难的.

4 Robocup 环境的 DID 模型及实验

动态影响图(DID)的决策过程是 Agent 通过对环境变化的推理和效用计算来选择具有最大期望效用的行为. 本文以 Robocup 的训练器为平台,利用 DID 对禁区内的 2 个球员配合射门问题进行建模.

设 A、B 球员为 2 个射门配合 Agents,C 为守门 Agent,当前球是被 A 控制. 在初始时间片中的 A、B 球员配合射门模型可用图 7 所示的影响图进行表示. B 和 C 结点各表示一个 Agent,其中每个 Agent 又对应一个如图 1 结构的影响图. D 为 A 和 B 的联合决策结点;U 为 A 和 B 的联合效用结点.

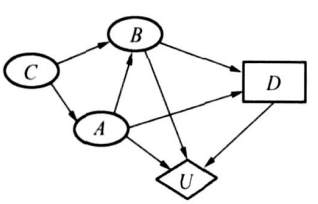


图7 初始影响图

Fig. 7 Original ID

根据射门球员所处位置对于射门的有利性,将图 7 所示的 Agent 连续的观察值空间离散成图 8 所示的网格区域,其中白圆、黑圆、空白和叉号分别表示有利、较有利、一般、不好的区域;状态结点离散成速度改变和方向改变的范围,其中方向可离散成周围 8 个网格对应的方向. A、B 球员的决策空间为(传球、跑位、射门),C 采用移向最强干扰区域的固定策略. 其中每个 Agent 的状态是不可观察的,而 Agent 所处的区域是可以观察的.

本文分别用 BK、粒子滤波(PF)以及联合树粒子滤波(JPF)[12]算法解决了 DID 模型的信度更新问题,并做了比较.图 9 是 A 球员在 50 个时间步内 3 种推理算法的信度概率误差(即 KL 距离).

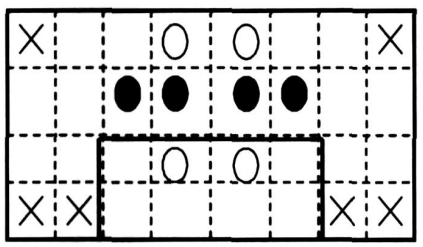


图 8　模型的训练场景

Fig. 8　Training scene of model

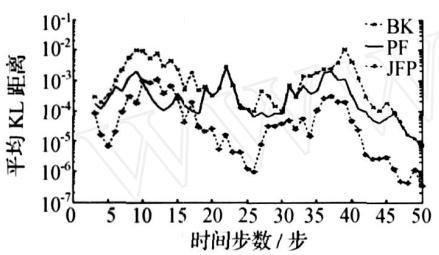


图 9　BK、PF 和 JFP 算法平均误差

Fig. 9　Average KL distance compared with BK, PF and JFP

利用足球机器人赛的训练器平台模拟比赛中的局部场景,用于学习局部协作.设 a_1、a_2、a_3、a_4 分别为 A 的有利、较有利、一般、不好的区域,则 A 的先验概率为

$$P(a_1) = 1/8,\quad P(a_2) = 1/8,$$
$$P(a_2) = 9/16,\quad P(a_4) = 3/16.$$

同样,设 b_1、b_2、b_3、b_4 分别为有利、较有利、一般、不好的区域.对于 B 有

$$P(b_1) = 1/8,\quad P(b_2) = 1/8,$$
$$P(b_3) = 9/16,\quad P(b_4) = 3/16.$$

C 对于 A 和 B 的整体干扰区域的先验概率为

$$P(c_1) = 1/3,\ P(c_2) = 1/3,\ P(c_3) = 1/3,$$

式中:c_1、c_2、c_3 分别表示为强、中和小干扰.

在进行 10 000 次训练中,以射门成功或失败作为训练结束标志,每次训练最多持续 20 个周期,如果在规定的周期内仍没有射门,则这次训练失败;以每 100 次训练中的进球数作为评价指标.实验结果如图 10 所示,图 10 表明当训练次数在 7 000 次时,成功率已接近 70%,相对于实际比赛而言,这已是一个很好的结果.

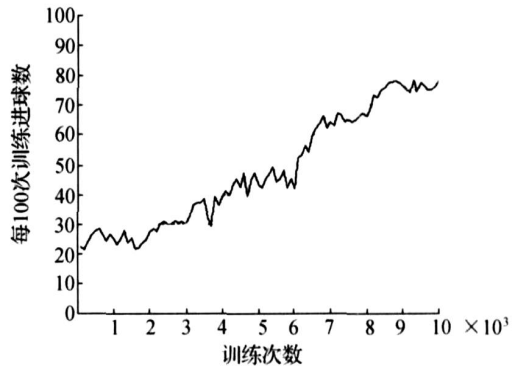


图 10　A 和 B 配合射门的成功次数

Fig. 10　Success times of score

5　结束语

DID 作为一个新的动态决策模型,相关问题还有待于进一步研究.把现有的 POMDP 求解算法应用于 DID 模型,并在实际问题中与 POMDP 比较是现在工作的一个主要方向.此外进一步把 DID 应用于多 Agent 系统中,以多 Agent 动态影响图来建模 Agent 之间的关系,实现多 Agent 之间的协作与决策,并与文献[13]中提出的多 Agent POMDP 模型 F-POMDP 模型也是一个有意义的工作.

参考文献:

[1] KAELBLING L P, LITTMAN M L, MOORE A W. Reinforcement learning: a survey[J]. Journal of Artificial Intelligence Research, 1996, 4: 237-285.

[2] POUPART P. Exploiting structure to efficiently solve large scale partially observable markov decision processes [D]. Toronto: University of Toronto, 2005.

[3] KAELBLING L P, LITTMAN M L, CASSANDRA A R. Planning and acting in partially observable stochastic domains[J]. Artificial Intelligence, 1998, 101: 99-134.

[4] MICHAEL J, YISHAY M, ANDREW Y. Ng approximate planning in large POMDPs via reusable trajectories [C]// Advances in Neural Information Processing Systems. [S.l.] Cambridge: MIT Press, 1999: 1001-1007.

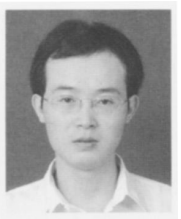
[5] NICHOLAS R, GEOFFREY J. Gordon, sebastian thrun: finding approximate POMDP solutions through belief compression[J]. J Artif Intell Res (JAIR), 2005, 23: 1-40.

[6] PAPADIMITRIOU C H, TSITSIKLIS J N. The complexity of Markov decision processes[J]. Mathematics of Operations Research, 1987, 12(3): 441-450.

[7] LUSENA C, GOLDSMITH J, MUNDHENK M. Non-approximability results for partially observable Markov

decision processes[J]. Journal of Artificial Intelligence Research, 2001, 14:83-103.
[8]DEAN T, KANAZAWA K. Probabilistic temporal reasoning[C]// National Conference on Artificial Intelligence. Washington:AAAI Press, 1988, 524-528.
[9]RONALD A, HOWARD, JAMES E. Readings on the principles and applications of decision analysis[M]. [S. l.]: Strategic Decision Group, 1984.
[10]BOUTILIER C, DEAN T, HANKS S. Decision-theoretic planning: structural assumptions and computational leverage[J]. Journal of Artificial Intelligence Research, 1999, 11:1-94.
[11]JOSEPH A, TATMAN, ROSS D, et al. Dynamic programming and influence diagrams[J]. IEEE Transations on Systems, Man, and Cybernetics, 1990, 20(2):365-379.
[12]BRENDA N G. Avi Pfeffer, Factored Particles for Scalable Monitoring[C]// Uncertainty in Artificial Intelligence Morgan Kaufmann. San Francisco, USA, 2002:370-377.
[13]PRASHANT D, PIOTR G. A particle filtering based approach to approximating interactive POMDPs[C]// P National Conference on Artificial Intelligence, Menlo Park. AAAI Press, 2005:969-974.

作者简介:



俞 奎,男,1979年生,硕士研究生,主要研究方向为贝叶斯网络建模与推理、Agent技术,发表学术论文7篇.

王 浩,男,1962年生,教授,博士,合肥工业大学计算机与信息学院副院长,主要研究方向为人工智能、数据挖掘、面向对象技术等,中国自动化学会机器人竞赛工作委员会委员、安徽省高校中青年骨干教师. 先后参加国家自然科学基金、国家教委博士点基金等10多项课题研究,获安徽省科技进步三等奖2项. 目前主持国家自然科学基金和安徽省自然科学基金等多项课题.

姚宏亮,男,1972年生,副教授,博士,主要研究方向为贝叶斯网络、Agent技术,发表学术论文10余篇.

2008 中国智能系统工程学术大会

2008 Conference on Intelligent systems Engineering

21世纪将是智能科学突飞猛进的世纪,将是智能系统工程大放异彩的世纪. 智能系统工程化、工程系统智能化将是人工智能走向社会、走向应用、走向成功的必由之路,将是人类科学开发智能、科学利用智能的必由之路. 当前,智能系统工程相关研究和应用呈现出方兴未艾的发展势头,研究队伍迅速扩大,研究领域急速拓展,一大批研究项目得到国家自然科学基金和国家科技攻关计划的支持,带动着信息科学向更高层次发展.

智能系统工程是人工智能的重要发展领域,为了推动和展示智能系统工程领域的研究进展,交流总结近年来智能系统工程领域的最新成果,中国人工智能学会智能系统工程专业委员会将于2008年在西南石油大学(中国 - 成都)组织召开中国智能系统工程学术大会,会议的主题是"智能系统工程的新发展". 本次会议将围绕智能系统工程特色创新成果进行交流与展示. 征文范围(但不限于): 1)理论创新;2)技术创新;3)应用创新;4)学科创新.

全文截稿日期: 2008-6-10.

会议开始日期: 2008-10-1.

会议网站: http://dxy.swpu.edu.cn/ReadNews.asp NewsID=515.