一种新颖的多 agent强化学习方法

周浦城, 洪炳镕, 黄庆成
(哈尔滨工业大学计算机科学与技术学院, 黑龙江哈尔滨 150001)

摘 要: 提出了一种综合了模块化结构、利益分配学习以及对手建模技术的多 agent强化学习方法, 利用模块化学习结构来克服状态空间的维数灾问题, 将 Q-学习与利益分配学习相结合以加快学习速度, 采用基于观察的对手建模来预测其他 agent的动作分布. 追捕问题的仿真结果验证了所提方法的有效性.

关键词: 多 agent学习; Q-学习; 利益分配学习; 模块化结构; 对手建模
中图分类号: TP18 文献标识码: A 文章编号: 0372-2112 (2006) 08-1488-04

A Novel Multi-Agent Reinforcement Learning Approach

ZHOU Pu-cheng, HONG Bing-rong, HUANG Qing-cheng
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract A novel multi-agent reinforcement learning approach is proposed to learn the coordinated behaviors among cooperative agents team. The proposed approach combines advantages of the modular architecture, profit-sharing learning and opponent modeling technique in a single multi-agent framework. Simulation results on the pursuit problem show that the proposed learning approach has faster convergence speed and more optimal policy over conventional modular Q-learning algorithms.

Key words multi-agent learning; Q-learning; profit-sharing learning; modular architecture; opponent modeling

1 引言

如何通过学习来协调各 agent之间的行为以适应动态变化的环境, 从而有效地实现共同的目标是多 agent系统研究的一个重要课题[1,2].

强化学习是一种以环境反馈作为输入的机械学习方法, 采用强化学习机制的 agent通过与环境进行交互来学习最优的行为策略[3]. 近年来强化学习逐渐成为多 agent系统的一类重要学习方法[4,5].

将强化学习技术应用到多 agent系统大致有两种方式, 一种方式是将多 agent系统作为单个学习的 agent, 并借助标准的单 agent强化学习技术来学习最优联合策略. 由于状态空间和动作空间将会随环境中 agent个数的增加呈指数倍增长, 这将导致维数灾问题. 另一种方式是让系统中每个 agent拥有各自独立的强化学习机制. 由于多个 agent同时在学习, 此时 agent的目标状态不仅取决于自己的行为, 同时还受环境中其他 agent行为效果的影响. 针对这些问题, 本文综合了模块化结构、利益分配学习以及对手建模等技术的优点, 在此基础上提出了一种新颖的多 agent强化学习方法, 并利用追捕问题进行了仿真验证.

2 强化学习

2.1 Q-学习

Q-学习是由 Watkins提出的一种类似于动态规划的强化学习方法[6]. 它提供 agent在马尔可夫环境中利用经历的动作序列来选择最优动作的学习能力, 而无需建立环境模型. Q-学习通过下面的更新规则来迭代逼近最优动作值函数:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left(r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right) \quad (1)$$

其中 $\alpha \in [0, 1]$是学习速率, $\gamma \in [0, 1]$是折扣系数, r_{t+1}是 agent执行动作 a_t 使得环境从状态 s_t 转移 s_{t+1}后所接收到的奖励值.

2.2 利益分配学习

利益分配 (Profit Sharing PS) 学习[7] 是强化学习算法的一种. 在 t时刻 agent处于状态 s_t, 按照一定的方式选择动作 a_t, 执行该动作后判断是否获得奖励 R. 若没有奖励, agent将称之为规则的状态-动作对 (state-action pair SAP) (s_t, a_t)存贮在事件记录中, 然后重复这一过程. 从初始状态到达最终奖励状态的过程称为一幕. 一旦 agent获得奖

收稿日期: 2005-12-01; 修回日期: 2006-03-15
基金项目: 国家自然科学基金 (No. 69985002)

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

励值R, 那么对事件记录中的规则进行如下更新:

$$w(s_t, a_t) \leftarrow w(s_t, a_t) + f(R, t) \quad (2)$$

其中 $w(s_t, a_t)$是 t时刻规则的权重, f是强化函数. 理性定理[8]保证那些构成回路的无效规则将会得到比有效规则更小的权重, 当 f满足:

$$f(R, t) > H \sum_{i=0}^{t-1} f(R, i) \quad (3)$$

其中H 是冲突的有效规则数的上界.

3 提出的模块强化学习方法

在合作的多 agent系统, 由于需要联合执行协作任务, 因此对于每个 agent其状态空间规模将随 agent的数量成指数增长, 导致状态空间组合爆炸. 为避免出现这一问题, 本文采用图 1所示的模块化学习结构. 学习 agent由三种模块构成: (1)学习模块 LM, 实现强化学习算法. (2)对手模块 OM, 用于通过观察其他 agent的动作来得到其动作分布估计, 以便评估自身动作值. (3)仲裁模块 MM, 用来综合学习模块和对手模块的结果.

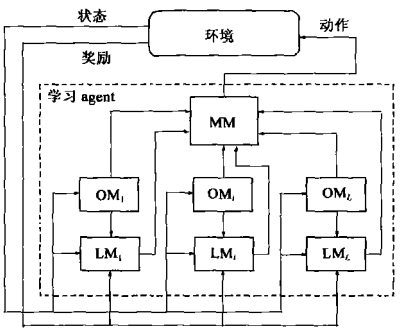

图 1 提出的模块化学习结构

3.1 对手建模

在多个 agent共存的环境中, agent的动作效果不仅取决于外界环境, 同时还受到其他 agent动作的影响. 为有效实现目标, agent应该考虑其他 agent的动作. 为此, 本文提出一种当 agent相互之间没有通讯的情况下基于观察的对手建模方法, 以便于各 agent预测其他 agent的行动.

在 t时刻, agent采取动作 a_s, 同时其他 agent执行动作 a_o. 环境由状态 s变为 s. 之后, agent辨别其他 agent在 t时刻执行的动作 a_o^*: 若在连续的两个时刻 t和 t+1, 其他 agent均能够被观察到, 那么 agent就可以根据其他 agent的位置变化来推断其采取的动作 a_o^*; 否则 a_o^* 是 UNCERTAIN. 如果 a_o^* 不是 UNCERTAIN, 那么对于状态 s下的所有可行动作 a_o, agent根据式(4)来更新相应的对手模型:

$$P_t(s, a_O) = \begin{cases} 1 - \dfrac{|A_O(s)| - 1}{|A_O(s)|} k, & a_O = a_O^* \\ k P_{t-1}(s, a_O), & \text{其他} \end{cases} \quad (4)$$

其中 $k \in [0, 1]$是确定先前动作分布效果的控制系数, $|A_O(s)|$是其他 agent在状态 s时可行动作的数量.

3.2 混合强化学习算法

研究表明 PS学习要比 Q学习更适合于动态或多 agent领域[9]. 然而对于 PS学习, 只有当到达了目标奖励状态后才更新 SAP的权重, 而不能在学习的过程中根据已有经验来获取最新知识. 为此, 本文提出将 Q学习与 PS学习相结合的混合强化学习方法. 在学习初期, 采用 Q学习算法来更新各 SAP的 Q值. 一旦到达奖励状态, 利用 PS学习来强化各 SAP. 令 $a_s \in A_s$ 以及 $a_o \in A_o$ 分别是 agent和其他 agent对应的动作, A_s 和 A_o 是各自相应的可行动作集. 各学习模块采用的混合强化学习算法描述如下.

(1)初始化. 对于所有的 $s \in S$, $a_s \in A_s(s)$, $a_o \in A_o(s)$:

$Q(s, a_s, a_o) \leftarrow$小的常数

(2)循环 (对于每一幕)

(a) $t \leftarrow 0$, 清空事件记录, 状态初始化

(b)循环(对于幕中每一步)

①观察当前状态 s_t

②根据仲裁模块选择的动作 a_s, 将 (s_t, a_s)保存到事件记录中

③agent执行动作 a_s, 其他 agent执行动作 a_o^*. 状态变成 s_{t+1}, 同时 agent获得奖励 r_{t+1}

④若新状态 s_{t+1}是奖励状态, $T \leftarrow t+1$, 转步骤③

⑤按如下方式更新 $Q(s_t, a_s, a_o)$:

(a)如果 $a_o \neq$UNCERTAIN

$$Q(s_t, a_s, a_o) \leftarrow (1-\alpha) Q(s_t, a_s, a_o) + \alpha [r_{t+1} + \gamma \max_{a' \in A_s(s_{t+1})} Q(s_{t+1}, a', a_o')]$$

其中 $a_O' = \arg\max_{b \in A_O(s_{t+1})} P(s_{t+1}, b)$.

(b)如果 $a_o =$ UNCERTAIN

对于所有的 $a_O \in A_O(s_t)$, 更新如下:

$$Q(s_t, a_s, a_o) \leftarrow (1-\alpha) Q(s_t, a_s, a_o) + \alpha [r_{t+1} + \gamma \max_{a' \in A_s(s_{t+1})} EQ(s_{t+1}, a')]$$

其中

$$EQ(s_{t+1}, a') = \sum_{b \in A_O(s_{t+1})} P(s_{t+1}, b) Q(s_{t+1}, a', b)$$

⑥$t \leftarrow t+1$, 转步骤①

(c)对于事件记录中所有 SAP, 按下式更新其 Q值:

$$Q(s_t, a_t, a_o) \leftarrow Q(s_t, a_t, a_0) + f(R, t) \quad (5)$$

其中R 是任务完成后得到的奖励. 为满足式(3)的要求, 这里采用如下函数:

$$f(R, t) = \gamma^{t-1} R \quad (6)$$

其中 $0 < \gamma \leqslant 1$, $\gamma = 1/\max|A(s)|$是折扣率.

3.3 仲裁模块决策

仲裁模块用于综合学习模块和对手模块的结果, 这里采用 GM (GreatestMass)策略来选择动作:

$$a_t = \arg\max_{a_s \in A_s} \sum_{k=1}^{L} EV^k(a_s) \quad (7)$$

其中 L是学习模块的数量, $EV^k(a_s)$是学习模块 k中动作 a_s 在考虑了对手模型后的期望动作值, 亦即:

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

$$EV^k(a_s) = \sum_{a_o \in A_o(s_t)} P(s_t, a_o) Q(s_t, a_s, a_o) \quad (8)$$

3.4 多 agent学习过程

①学习模块和对手模块观察当前状态 s_t.

②对手模块将其他 agent相应的动作分布值传给仲裁模块, 同时学习模块也将 Q值传给仲裁模块.

③仲裁模块根据 GM 策略选择动作 a_s. 学习模块将 (s_t, a_s)保存在事件记录中.

④agent执行动作 a_s, 其他 agent执行动作 a_o. 环境转移到新的状态 s_{t+1}, agent获得奖励值 r_{t+1}.

⑤agent判别其他 agent刚才执行的动作 a_o; 并更新其对手模型 $P(s_t, a_o')$.

⑥若状态 s_{t+1}是终止状态, 那么转步骤⑨.

⑦agent更新 Q-函数, $Q(s_t, a_s, a_o')$

⑧$t \leftarrow t+1$ 转步骤①.

⑨若状态 s_{t+1}是奖励状态, 那么利用 PS学习分配得到的奖励值 R.

⑩若算法满足结束条件, 学习结束, 否则转步骤①.

4 仿真

为测试本文提出的算法, 利用追捕问题进行了仿真, 并分析了算法的运行效果.

4.1 追捕问题

在如图 2(a)所示 10×10的栅格模型中, 4个猎人 agent(简称 Hunter)围捕 1个随机运动的猎物 agent(简称 Prey), Hunter是学习 agent. 在每一时刻, agent有 5种候选动作: 上移, 下移, 左移, 右移及留在原地. agent每次移动一个栅格的位置, 允许数个 Hunter共处一个栅格, 但不能与 Prey处在同一个栅格, 且不允许 agent穿越边界. 视野半径为 d的 agent能够看见大小为 $v_s = (2l+1) \times (2l+1)$的区域内所有的栅格. 每个 agent分配了唯一的 ID, 各 agent能够识别视野范围内的其他 agent并获悉其相对于自己的位置信息. 各 agent独立决策, 相互之间不允许通讯. 当 Hunter围住 Prey致使其不能移动时围捕成功, 图 2(b)是围捕成功的一些典型情形.

4.2 实验结果

在 Q-学习中采用的仿真参数是: 学习速率 $\alpha = 0.8$, 折扣系数 $\gamma = 0.9$, 各 SAP的初始 Q值设为 0.1, 各对手模型函数的初始值设为 0.2. 每次试验时总的学习幕数为 5000, 限定每一幕的最大迭代步长为 1500步, 若在此期间未能围捕成功则放弃本次学习, 重新开始学习. 一旦捕获 Prey, 所有 Hunter均得到奖励值 1.0, 其他情况下奖励值为 -0.1.

设定 Hunter的视野半径 d=3. 每个学习模块在状态表达时仅考虑 Prey和特定的队友 agent相对自己的位置, 此时状态空间规模 $|S| = ((2l+1)^2+1)^2 = 2500$, 从而 Hunter总的状态空间规模为 7500. 对于采用查找表的传统 Q-学习, 当状态表达为自己与所有其他 agent相对位置的组合, 那么此时状态空间规模大约为 $((2l+1)^2+1)^4 = 6250000$ 显然后者需要更多的存贮资源.

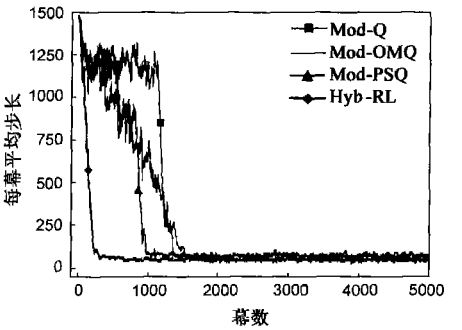

图 3 不同学习算法的比较

首先比较了本文提出的方法(记为 Hyb-RL)与模块 Q-学习(记为 Mod-Q)[10]、基于对手建模的模块 Q-学习(记为 Mod-OMQ)以及基于 PS学习的模块 Q-学习(记为 Mod-PSQ), 这里的 Mod-OMQ 算法源于 Kaya等人的研究工作[11], 但未采用模糊逻辑. PS学习中折扣率 $\lambda = 0.2$, 对手建模的控制系数 $k = 0.8$. 图 3给出了 4种学习算法得到的学习曲线. 可以看出: (1)综合了 PS学习的 Q-学习能够显著加快学习速度. 实验表明, 与 Mod-Q以及 Mod-OMQ 相比, 本文提出的方法收敛速度大约提高了 4至 5倍. (2)提出的方法得到的策略要优于其他学习算法.

为评价本文提出的对手建模方法(记为 OBM), 利用基于虚拟行动的对手建模[12](记为 FP)以及 Nagayuki提出的内部模型[13](记为 IM)进行了对比试验. 图 4给出了 1000幕得到的学习结果. 可以看出, 本文给出的对手建模

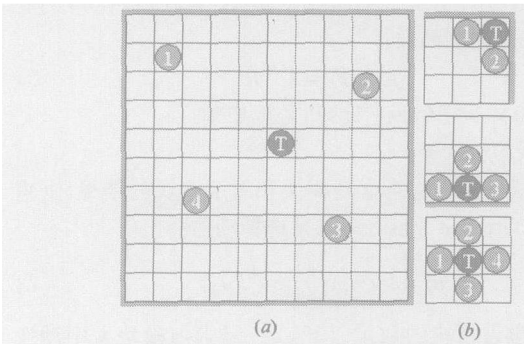

图 2 追捕问题. (a) 初始状态. (b) 围捕成功示例

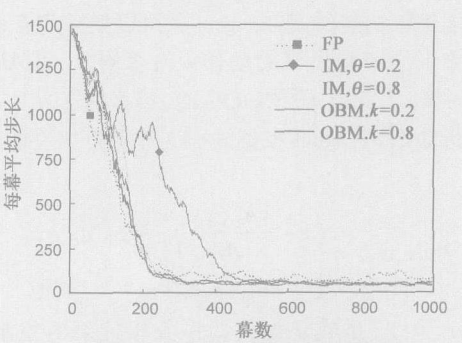

图 4 不同对手建模方法的比较

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

方法受参数变化的影响较小, 因此比内部模型更鲁棒. 另外, 虽然基于虚拟行动的对手建模与本文方法有相近的收敛速度, 但得到的策略要差一些.

最后, 测试了不同 PS学习参数对本文提出的学习方法的影响, 结果如图 5所示. 为了清晰起见, 这里仅给出前 1000幕的结果, 图中每个采样点对应于 50幕的平均学习结果. 从总的趋势来看, 随着 PS学习中折扣率的增长, 提出的学习方法收敛速度逐渐变快.

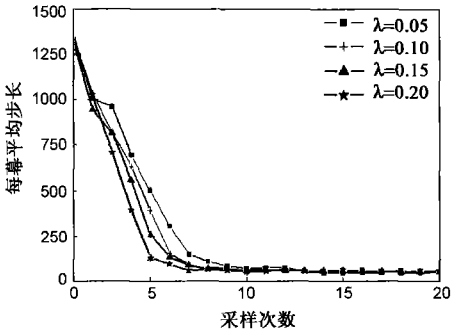

图 5　提出的算法在不同参数下的学习结果

5 结论

如何协调各 agent的行为是多 agent系统研究的一个重要课题. 针对合作的 agent群体在无通讯条件下的协作问题, 本文提出了一种综合了模块化结构、利益分配学习以及对手建模技术的多 agent强化学习方法. 提出的方法采用模块化学习结构来克服状态空间的维数灾问题, 并将 Q学习与利益分配学习相结合来加快学习速度, 与此同时利用基于观察的对手建模来预测其他 agent的动作分布, 以便更准确地估计动作值函数. 通过追捕问题的仿真结果表明, 提出的方法要比传统的模块 Q学习算法收敛速度更快, 得到的策略更优.

参考文献:

[1] Ho F, Kamel M. Learning coordinating strategies for cooperative multiagent systems[J]. Machine Learning, 1998, 33(2-3): 155- 177.

[2] Garland A, Alterman R. Autonomous agents that learn to better coordinate[J]. Autonomous Agents and MultiAgent Systems, 2004, 8(3): 267- 301.

[3] Kaelbing L P, Littman M L, Moore A W. Reinforcement learning A survey[J]. Journal of Artificial Intelligence Research, 1996, 4: 237- 285.

[4] Brafman R I, Tennenholtz M. Learning to coordinate efficiently: A model-based approach[J]. Journal of Artificial Intelligence Research, 2003, 18: 517- 529.

[5] Chen G, Yang Zh. Coordinating multiple agents via reinforcement learning[J]. Autonomous Agents and MultiAgent Systems, 2005, 10(3): 273- 328.

[6] Watkins C J C H, Dayan P. Technical note: Q-learning[J]. Machine learning, 1992, 8(3-4): 279- 292.

[7] Grefenstette J J. Credit assignment in rule discovery systems based on genetic algorithms[J]. Machine Learning, 1988, 3(2-3): 225- 245.

[8] Miyazaki K, Kobayashi S. Proposal for an algorithm to improve a rational policy in POMDPs[A]. Proc of IEEE International Conference on SMC[C]. Japan, Tokyo: IEEE SMC Society, 1999. 492- 497.

[9] Arai S, Sycara K. Effective learning approach for planning and scheduling in multi-agent domain[A]. Proc of the 6th ISAB[C]. France, Paris: The MIT Press, 2000. 507- 516.

[10] Ono N, Fukumoto K. Multi-agent reinforcement learning: A modular approach[A]. Proc of the 2th ICMAS[C]. USA, Portland: AAAI Press, 1996. 252- 258.

[11] Kaya M, Alhajj R. Modular fuzzy-reinforcement learning approach with internal model capabilities for multiagent systems[J]. IEEE Trans SMC-B, 2004, 34(2): 1210- 1223.

[12] Claus C, Boutilier C. The dynamics of reinforcement learning in cooperative mutliagent systems[A]. Proc of the 15th National Conference on Artificial Intelligence[C]. USA, Menlo Park: AAAI Press, 1998. 746- 752.

[13] Nagayuki Y, Ishii S, Doya K. Multi-agent reinforcement learning: An approach based on the other agent's internal model[A]. Proc of IEEE ICMAS[C]. USA, Boston: IEEE Computer Society, 2000. 215- 221.

作者简介:

周浦城　男, 1977年生于江西宜春, 分别于 1999年和 2002年获解放军炮兵学院军事指挥专业学士学位和军事运筹专业硕士学位, 现为哈尔滨工业大学计算机科学与技术学院计算机应用技术专业博士研究生, 主要从事多 agent系统、强化学习、智能机器人等方面的研究.
E-mail: zhoupc@hit.edu.cn

洪炳镕　男, 教授, 博士生导师, 1937年生于吉林延边, 1983年获日本早稻田大学工学博士学位. 现为哈尔滨工业大学计算机科学与技术学院教授, 中国人工智能学会常务理事, 国际机器人足球联盟副主席. 主持过多项国家 863计划、国家自然科学基金、省部委和国际合作科研项目. 成果获省部级科技进步一等奖 1项、二等奖 4项、三等奖 6项, 在国内外学术刊物上发表论文 200多篇, 著书 4部. 主要研究领域为: 分布式人工智能、智能机器人、多机器人系统、虚拟现实等.

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