

基于 MAXQ 分层分解的马尔科夫决策问题 在线规划算法研究

姓名： 柏爱俊

学号： BA11011028

导师： 陈小平

计算机科学与技术学院
中国科学技术大学

2016 年 6 月 7 日

主要内容

- ① 问题背景
- ② 研究现状
- ③ 研究内容
- ④ 进度安排

RoboCup 仿真 2D 机器人足球

- 11 对 11 足球比赛
- 完全分布式自主决策
- 带噪音的局部观察
- 带随机误差的原子动作
- 有限带宽的通讯
- 合作与对抗
- 实时系统 (100ms/周期)

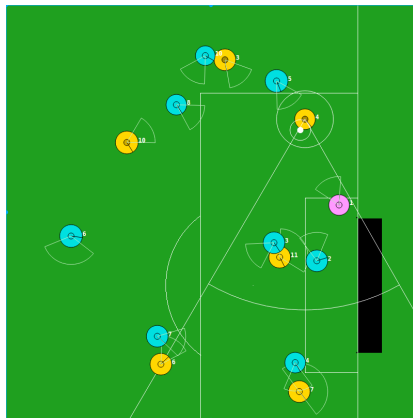


图 1: RoboCup 2D

多目标多传感器系统管理和调度

- 开放空间
- 目标随时出现或消失
- 目标运动轨迹不可预测
- 完全分布式自主决策
- 受限通讯
- 合作完成发现、跟踪等任务

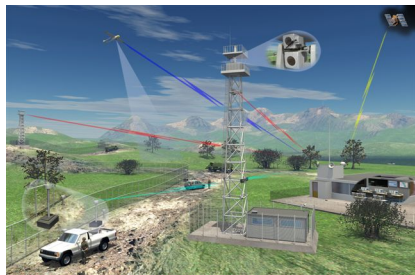


图 2: 传感器管理

不确定环境下多智能体合作与对抗在线规划

- 主要挑战：
 - 完全分布式
 - 局部观察
 - 受限通讯
 - 多智能体系统
 - 动作不确定性
 - 观察不确定性
 - 问题规模巨大：
 - 状态空间
 - 动作空间
 - 观察空间

马尔科夫决策理论

- 马尔科夫决策理论 (Markov Decision Theory) 为不确定性环境下的规划问题提供了基本的理论框架 [Puterman, 1994]
- 马尔科夫性：系统的状态转移仅依赖于当前状态和智能体执行的动作，不依赖于历史状态和其他信息
- 根据具体环境的不同，一般细分为 3 类问题：
 - ① 单智能体完全可观察——马尔科夫决策过程 (MDP)
 - ② 单智能体部分可观察——部分可观察 MDP (POMDP)
 - ③ 多智能体部分可观察——分布式 POMDP (DEC-POMDP)

马尔科夫决策过程

- MDP 仅考虑动作不确定性：

- ① 状态空间： $S = \{s_1, s_2, \dots, s_{|S|}\}$
- ② 动作空间： $A = \{a_1, a_2, \dots, a_{|A|}\}$
- ③ 转移函数： $T(s'|s, a) \rightarrow [0, 1]$
- ④ 收益函数： $R(s, a) \rightarrow \mathbb{R}$
- ⑤ 求解策略： $\pi : S \rightarrow A$

- POMDP 同时考虑观察不确定性：

- ① 观察空间： $O = \{o_1, o_2, \dots, o_{|O|}\}$
- ② 观察函数： $Z(o|a, s') \rightarrow [0, 1]$
- ③ 求解策略： $\pi : \Delta(S) \rightarrow A$

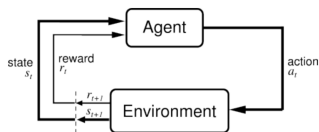


图 3: MDP 问题

马尔科夫决策过程（续）

- DEC-POMDP 是 POMDP 在多智能体系统中的自然扩展：

- ① 智能体集合： I
- ② 联合动作空间： $\vec{A} = \times_{i \in I} A_i$
- ③ 联合观察空间： $\vec{O} = \times_{o \in I} O_i$
- ④ 联合转移、观察和收益函数
- ⑤ 求解策略： $\pi_i : H_i \rightarrow A_i$

问题复杂度

- $P \subseteq NP \subseteq PSPACE \subseteq EXP \subseteq NEXP$
- MDP : P
- POMDP : PSPACE
- DEC-POMDP : NEXP

维度诅咒

状态空间随状态维度成指数式增长

研究现状

- 离线算法：

- 遍历整个状态空间，可以精确求解出最优策略
- 在大规模问题中离线算法不可用
- 线性规划、值迭代、策略迭代等 [Puterman, 1994]

- 在线算法：

- 从当前状态出发向前搜索，无需遍历整个状态空间
- 但如不利用具体问题的内部结构，允许的搜索深度会很有限
- 实时动态规划、与或图搜索、蒙特卡罗树搜索等 [Ross et al., 2008]

- 分层分解：

- 利用问题的内部结构，将其分解成规模较小的子问题
- 宏状态、宏动作、分层强化学习等 [Barto and Mahadevan, 2003]

基于 MAXQ 分层分解的在线规划算法

- 研究内容：开发基于 MAXQ **分层分解** 的**在线规划**算法，求解大规模不确定环境下多智能体合作与对抗问题
- MAXQ 分层方法来源于分层强化学习 [Dietterich, 1999]，利用层次结构，将一个 MDP 分解成一系列子问题：
 - $M = \{M_0, M_1, \dots, M_n\}$
 - $M_i = \{T_i, A_i, R_i\}$ ：
 - 终止条件 T_i
 - 动作空间 A_i （宏动作或原子动作）
 - 局部收益函数 R_i
 - 求解 M_0 即求解了原始问题 M

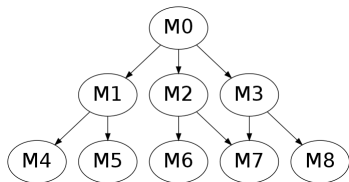


图 4: MAXQ 任务图

研究方案

● 拟进行的技术路线：

- 多智能体环境中，假设其他智能体具有有限的策略集合
- 假设其他智能体会根据世界状态的变化来选择自己的策略
- 把其他智能体实时选择的策略跟世界状态一起组成新的联合状态
- 使用 POMDP[Cassandra et al., 1995] 完整建模这一问题
- 使用 MAXQ[Dietterich, 1999] 方法分层分解建模后的 POMDP 问题
- 使用粒子滤波表达和更新信念状态 [Silver and Veness, 2010]
- 使用蒙特卡罗树搜索 [Browne et al., 2012] 在线求解
- 在此基础上，提出有效的在线通讯策略 [Wu et al., 2011]

研究方案（续）

● 相关工作：

- [Barry et al., 2011] 提出了大规模 MDP 问题分层求解方法，但其假设宏状态之间具有确定的转移路径，很多实际问题并不适用
- [Gmytrasiewicz and Doshi, 2005] 通过假设其他智能体的意图模型，提出了适用于多智能体系统的交互式 POMDP 模型，但由于其不可避免的信念嵌套问题，实际无法计算，从而只具有理论意义
- [Ramírez and Geffner, 2011] 假设其他智能体具有确定的目标（Goal），使用 POMDP 建模并求解了目标识别问题，而很多实际问题中智能体的目标是可能会随时间变化的

● 本工作创新点：

- 基于 MAXQ 分层分解的在线规划算法，可处理所有层次的不确定性
- 假设其他智能体具有有限的策略集合，避免了信念嵌套的问题
- 其他智能体选择的策略不是固定不变的，可适用于更多的实际问题
- 结合粒子滤波和蒙特卡罗树搜索，实现高效在线求解算法

研究方案（续）

- 实验平台：

- RoboCup 2D 中的合作与对抗问题
- 多目标多传感器系统管理和调度

- 预期成果：

- 在线近似求解目前文献上尚无法求解的大规模不确定环境下多智能体合作与对抗问题
- 应用到科大“蓝鹰”仿真 2D 机器人足球队，争取在 RoboCup 2D 世界杯比赛中取得佳绩
- 在本领域国际知名学术出版物和重要学术会议发表论文 3-4 篇

已有工作

- MAXQ-OP：基于 MAXQ 分层分解的 MDP 在线规划算法
[Bai et al., 2012b]
- 主要贡献：
 - 提出了利用 MAXQ 分层结构的在线规划算法框架
 - 完成了对 MAXQ 分层结构中完成函数的在线近似计算
- 实验平台：
 - 出租车问题（标准测试问题）[Bai et al., 2012c]
 - RoboCup 2D 中的行为分层和在线规划
[Bai et al., 2012a, Bai et al., 2013, Bai et al., 2012c]

已发表论文

- Bai, A., Chen, X., MacAlpine, P., Urieli, D., Barrett, S. and Stone, P. (2012a).
Wright Eagle and UT Austin Villa: RoboCup 2011 Simulation League Champions.
In RoboCup-2011: Robot Soccer World Cup XV, (Roefer, T., Mayer, N. M., Savage, J. and Saranli, U., eds), vol. 7416, of Lecture Notes in Artificial Intelligence. Springer Verlag Berlin
- Bai, A., Wu, F. and Chen, X. (2012b).
Online Planning for Large MDPs with MAXQ Decomposition (Extended Abstract).
In Proc. of 11th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2012)
- Bai, A., Wu, F. and Chen, X. (2012c).
Online Planning for Large MDPs with MAXQ Decomposition.
In Proc. of the Autonomous Robots and Multirobot Systems workshop (at AAMAS-12)
- Bai, A., Wu, F. and Chen, X. (2013).
Towards a Principled Solution to Simulated Robot Soccer.
In RoboCup-2012: Robot Soccer World Cup XVI, (Chen, X., Stone, P., Sucar, L. E. and der Zant, T. V., eds), vol. 7500, of Lecture Notes in Artificial Intelligence. Springer Verlag Berlin

研究进度安排

- ① 2013.1 - 2013.3：调研本领域研究前沿和热点问题，确定基本假设
- ② 2013.4 - 2013.7：给出假设条件下适用于多智能体系统的算法框架
- ③ 2013.8 - 2013.11：实验验证，并持续改进算法
- ④ 2013.12 - 2014.2：理论分析，并跟相关工作进行全面比较
- ⑤ 2014.3 - 2014.6：整理最终研究成果，撰写论文

参考文献 I



Bai, A., Chen, X., MacAlpine, P., Urieli, D., Barrett, S. and Stone, P. (2012a).
Wright Eagle and UT Austin Villa: RoboCup 2011 Simulation League Champions.
In *RoboCup-2011: Robot Soccer World Cup XV*, (Roefer, T., Mayer, N. M., Savage, J. and Saranli, U., eds), vol. 7416, of
Lecture Notes in Artificial Intelligence. Springer Verlag Berlin.



Bai, A., Wu, F. and Chen, X. (2012b).
Online Planning for Large MDPs with MAXQ Decomposition (Extended Abstract).
In *Proc. of 11th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2012)*.



Bai, A., Wu, F. and Chen, X. (2012c).
Online Planning for Large MDPs with MAXQ Decomposition.
In *Proc. of the Autonomous Robots and Multirobot Systems workshop (at AAMAS-12)*.



Bai, A., Wu, F. and Chen, X. (2013).
Towards a Principled Solution to Simulated Robot Soccer.
In *RoboCup-2012: Robot Soccer World Cup XVI*, (Chen, X., Stone, P., Sucar, L. E. and der Zant, T. V., eds), vol. 7500,
of *Lecture Notes in Artificial Intelligence*. Springer Verlag Berlin.



Barry, J., Kaelbling, L. and Lozano-Perez, T. (2011).
DetH*: Approximate Hierarchical Solution of Large Markov Decision Processes.
In *International Joint Conference on Artificial Intelligence pp. 1928–1935*.



Barto, A. and Mahadevan, S. (2003).
Recent advances in hierarchical reinforcement learning.
Discrete Event Dynamic Systems 13, 341–379.

参考文献 II



Browne, C., Powley, E. J., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S. and Colton, S. (2012).

A Survey of Monte Carlo Tree Search Methods.

IEEE Trans. Comput. Intellig. and AI in Games 4, 1–43.



Cassandra, A., Kaelbling, L. and Littman, M. (1995).

Acting optimally in partially observable stochastic domains.

In Proceedings of the National Conference on Artificial Intelligence pp. 1023–1023, JOHN WILEY & SONS LTD.



Dietterich, T. G. (1999).

Hierarchical Reinforcement Learning with the MAXQ Value Function Decomposition.

Journal of Machine Learning Research 13, 63.



Gmytrasiewicz, P. and Doshi, P. (2005).

A framework for sequential planning in multiagent settings.

Journal of Artificial Intelligence Research 24, 49–79.



Puterman, M. L. (1994).

Markov Decision Processes: Discrete Stochastic Dynamic Programming.

John Wiley & Sons, Inc.






Ramírez, M. and Geffner, H. (2011).

Goal recognition over POMDPs: Inferring the intention of a POMDP agent.

In Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three pp. 2009–2014, AAAI Press.

参考文献 III

- 
- Ross, S., Pineau, J., Paquet, S. and Chaib-Draa, B. (2008).
Online planning algorithms for POMDPs.
[Journal of Artificial Intelligence Research](#) 32, 663–704.
- 
- Silver, D. and Veness, J. (2010).
Monte-Carlo planning in large POMDPs.
[Advances in Neural Information Processing Systems \(NIPS\)](#) 46.
- 
- Wu, F., Zilberstein, S. and Chen, X. (2011).
Online planning for multi-agent systems with bounded communication.
[Artificial Intelligence](#) 175, 487–511.