第 28 卷 第 3 期
2002 年 5 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 28, No. 3
May, 2002

短文

基于一般和随机对策论框架下的多智能体学习[1)]

欧海涛　　张卫东　　许晓鸣

（上海交通大学自动化系　上海　200030）

（E-mail：haitaoou@yahoo.com）

摘　要　将 Q-learning 从单智能体框架上扩展到非合作的多智能体框架上，建立了在一般和随机对策框架下的多智能体理论框架和学习算法，提出了以 Nash 平衡点作为学习目标. 给出了对策结构的约束条件，并证明了在此约束条件下算法的收敛性，对多智能体系统的研究与应用有重要意义.

关键词　多智能体，Q-learning，随机对策，Nash 平衡点

中图分类号　TP13

MULTI-AGENT LEARNING BASED ON GENERAL-SUM STOCHASTIC GAMES

OU Hai-Tao　ZHANG Wei-Dong　XU Xiao-Ming

(*Department of Automation*, *Shanghai Jiaotong University*, *Shanghai*　200030)

(E-mail: haitaoou@yahoo.com)

Abstract　Q-learning from original single-agent framework is extended to non-cooperative multi-agent framework, and the theoretic framework of multi-agent learning is proposed under general-sum stochastic games with Nash equilibrium point as learning objective. We introduce a multi-agent Q-learning algorithm and prove its convergence under certain restriction, which is very important for the study and application of multi-agent system.

Key words　Multi-agent, Q-learning, stochastic games, Nash equilibrium point

1　引言

强化学习是一种无模型并可在线进行学习的方法，特别是对于动态环境变化和智能体

1) 国家自然科学基金(60174038)资助

收稿日期　2000-01-14　　收修改稿日期　2001-01-11

之间的不完全信息等特征,非常适合于多智能体系统[1]. Littman[2,3]改进了单个智能体的强化学习方法使之更适合于多智能体情况.他提出针对零和随机对策的最小最大 Q-learning 方法,但仅仅局限于零和对策或重复对策两种情况.本文研究了在一般和随机对策(又称马尔可夫对策)框架下的强化学习,主要是非合作系统的多智能体学习,对于合作系统的学习由于多智能体间可以通讯和作出承诺而大不相同.

2 对多智能体系统建模

一个多智能体系统的模型和单智能体系统的模型最大区别就是智能体直接认识到其它智能体的存在并对其行为进行建模.我们主要研究的是非合作多智能体系统,智能体之间不存在联合行动的协议[4],其中满足 Markov 状态转移的非合作对策就称为随机对策.

2.1 Nash 平衡点

Nash 平衡点表示对策中局中人对其它局中人合理性行为动作的正确估计,从而达到一种稳定状态.合理性动作意味着每一个智能体的策略都是对其它智能体既定策略的最优响应[5].这个概念是 Nash 在 1951 年提出的,已经广泛应用作为解决一般和非合作对策的主要方法.

2.2 多智能体 Q-learning

在多智能体环境中,一个智能体有两种方法对其它智能体建模.一种是忽略它们的个性,将其作为环境的一部分加以考虑;另一种就是将其清晰地看作是理性决策个体.两种方法的区别是在建模的难易程度、计算复杂性和预测能力上.我们的研究主要是针对不完全信息情况下使用 Q-learning 求解最优行动,作为一种计算方法 Q-learning 求解 Nash 平衡点不需要知道转移概率的知识;另一方面在不完全信息情况下学习期间 Q 值提供到最优值的最佳逼近.应用 Q-learning 主要有两个关键问题:确定学习函数和如何更新 Q 函数.我们在随机对策中定义学习函数,在学习函数中体现联合行动意图,并定义 Nash 平衡点求取作为学习目标.采用 Nash 平衡点作为求解目标的合理性是基于两点:1)我们假设所有的智能体都是理性的,都对其它智能体采取优化响应;2)Nash 平衡点表示在它们具有的信息基础上智能体相互合理作用的一种长期稳定状态.

2.2.1 多智能体 Q 函数

对于一个 n 个局中人的随机对策,定义智能体 k 的 Nash 平衡点 Q 值为

$$Q_*^k(s,a^1,\cdots,a^n)=r^k(s,a^1,\cdots,a^n)+\beta\sum_{s'\in s}p(s'\,|\,s,a^1,\cdots,a^n)v^k(s',\pi_*^1,\cdots,\pi_*^n) \quad (1)$$

Nash 平衡点 Q 值定义为在状态时所有智能体执行联合行动$(a^1,\cdots,a^n)$,并遵照 Nash 平衡点策略所得到的报酬.

2.2.2 一种多智能体 Q-learning 算法

多智能体 Q-learning 算法和一般 Q-learning 算法的区别主要在于:1)我们学习的 Q 函数是所有智能体的联合行动的函数,而一般 Q-learning 中则仅仅是一个智能体行动的函数;2)我们算法中的 Q 函数的更新是假设在智能体的优化决策都是 Nash 平衡点行动基础上的,而一般 Q-learning 中是在对其自身 Q 值最大的基础上的优化选取来更新的.

一个 Q 值表可以分解成一系列子表,即 $Q^k=(Q^k(s^1),\cdots,Q^k(s^m))$是智能体 k 的 Q 值表.

$Q^k(s^i)$是在状态 s^j 时的 Q 值表,表示 $Q^k(s^j,a^1,\cdots,a^n)$. $Q^k(s^j)$的全部项数是 $\prod_{i=1}^{n}|A^i|$. 智能体 k 根据下面的法则更新 Q 值:

$$Q_{t+1}^k(s,a^1,\cdots,a^n)=(1-\alpha_t)Q_t^k(s,a^1,\cdots,a^n)+\alpha_t[r_t^k+\beta\pi^1(s_{t+1})\cdots\pi^n(s_{t+1})Q_t^k(s_{t+1})],$$

其中 $(\pi^1(s_{t+1}),\cdots,\pi^n(s_{t+1}))$ 是对正规形对策$(Q_t^1(s_{t+1}),\cdots,Q_t^n(s_{t+1}))$ 和 $\alpha_t=0$ 时$(s,a^1,\cdots,a^n)\neq(s_t,a_t^1,\cdots,a_t^n)$ 的混合策略 Nash 平衡点. 这里有两点说明: 1) $\pi^1(s_{t+1})\cdots\pi^n(s_{t+1})Q_t^k(s_{t+1})$ 是级 数,表示智能体 k 在状态 s_{t+1}处的期望 Q 值; 2)上式并不更新 Q 值表中所有项目,只是更新与当前状态及智能体所选行动对应的项目. 在对策开始时,智能体不具备除其它智能体行为空间外的任何信息,随着对策进行,智能体 k 观测其它智能体的即时报酬和以前的行动,将此信息用来更新智能体 k 对其它智能体 Q 值表的推断. 智能体 k 更新其关于智能体 j 的 Q 值表的信念,根据如下法则进行:$j\neq k$,

$$Q_{t+1}^j(s,a^1,\cdots,a^n)=(1-\alpha_t)Q_t^j(s,a^1,\cdots,a^n)+\alpha_t[r_t^j+\beta\pi^1(s_{t+1})\cdots\pi^n(s_{t+1})Q_t^j(s_{t+1})].$$

下面给出学习算法的基本步骤:

1) 初始化. $t=0$,对所有的 $s\in S$, $a^k\in A^k$, $k=1,\cdots,n$, 使 $Q_t^k(s,a^1,\cdots,a^n)=0$, 初始化状态 s_0, 给一初值;

2) LOOP. 选取行动 a_t^i, 观测 $r_t^1,\cdots,r_t^n$; $a_t^1,\cdots,a_t^n$ 和 s_{t+1}, 更新 Q^k, $k=1,\cdots,n$,

$$Q_{t+1}^k(s,a^1,\cdots,a^n)=(1-\alpha_t)Q_t^k(s,a^1,\cdots,a^n)+\alpha_t[r_t^k+\beta\pi^1(s_{t+1}),\cdots,\pi^n(s_{t+1})Q_t^k(s_{t+1})],$$

其中($\pi^1(s_{t+1}),\cdots,\pi^n(s_{t+1})$) 是对正规形对策$(Q_t^1(s_{t+1}),\cdots,Q_t^n(s_{t+1}))$ 的混合策略 Nash 平衡点, $t:=t+1$.

3　学习算法的收敛性证明

本文只对二人随机对策的 Q-learning 算法给出收敛性证明. 同理,结论可以推广到 n 人随机对策也是成立的. 首先给出 Q-learning 的一般假设.

假设 1. 每一个状态和行动都被算法无穷遍历到.

假设 2. 学习速率满足下列条件:

1) $0\leqslant\alpha_t\leqslant1$, $\sum_{t=0}^{\infty}\alpha_t=\infty$ and $\sum_{t=0}^{\infty}\alpha_t^2<\infty$;

2) $\alpha_t(s,a^1,a^2)=0$ if$(s,a^1,a^2)\neq(s_t,a_t^1,a_t^2)$.

算法收敛性证明中需要用到的以下引理是由 Littman 证明得到.

引理 1. 在假设 1 和 2 的情况下,$Q_{t+1}=(1-\alpha_t)Q_t+\alpha_t\omega_t$ 收敛于 $E(\omega_t|h_t,\alpha_t)$,其中 h_t 是历史记录.

引理 2. 在假设 1 和 2 的情况下, $U_{t+1}(x)=(1-\alpha_t)U_t(x)+\alpha_t[P_tv^*](x)$收敛于 v^*,并且 P_t 对所有的 V 满足 $\|P_tV-P_tv^*\|\leqslant\gamma\|V-v^*\|+\lambda_t$,其中 $0<\gamma<1$ 和 $\lambda_t\geqslant0$ 收敛于 0, 则 $V_{t+1}(x)=(1-\alpha_t)V_t(x)+\alpha_t[P_tV_t](x)$收敛于 v^*.

定理 1. 下面的两种表达是等价的(由 Filar 和 Vrieze 证明得到[6]).

1) 对每一个 $s\in S$,对$(\pi^1(s),\pi^2(s))$组成了静态双矩阵对策$(Q^1(s),Q^2(s))$的一个平衡点,平衡点支付为$(v^1(s,\pi^1,\pi^2),\ v^2(s,\pi^1,\pi^2))$,对 $k=1,2$,有

$$Q^k(s,a^1,a^2)=r^k(s,a^1,a^2)+\beta\sum_{s\in S}p(s'|s,a^1,a^2)v^k(s',\pi^1,\pi^2);$$

2) (π^1,π^2)是在折扣随机对策 Γ 中的平衡点,平衡点支付为$(v^1(\pi^1,\pi^2),v^2(\pi^1,\pi^2))$,其中 $v^k(\pi^1,\pi^2)=(v^k(s^1,\pi^1,\pi^2),\cdots,v^k(s^m,\pi^1,\pi^2))$, $k=1,2$.

我们证明了多智能体 Q-learning 算法收敛于 Nash 平衡点 Q 值的以下引理和定理,限于篇幅的原因,这里只给出了它们的描述.

引理 3. 使 $P_tQ=(P_t^1Q^1,P_t^2Q^2)$满足 $P_t^1Q^1(s,a^1,a^2)=r_t^1+\beta\pi^1(s_t)Q^1(s_t)Q^1(s_t)\pi^2(s_t)$ 和 $P_t^2Q^2(s,a^1,a^2)=r_t^2+\beta\pi^1(s_t)Q^2(s_t)\pi^2(s_t)$, 其中$(\pi^1(s_t),\pi^2(s_t))$ 是对于双矩阵策略 $(Q^1(s_t),Q^2(s_t))$ 的混合策略 Nash 平衡点,对于所有 Q: $\|P_tQ-P_tQ_*\|\leqslant\beta\|Q-Q_*\|$,其中$Q_*=(Q_*^1,Q_*^2)$的元素由(2)式定义. 证明略.

定理 2. 在随机对策 Γ 中,在假设 1 和 2 的情况下,序列对$\{Q_t^1,Q_t^2\}$由式

$Q_{t+1}^k(s,a^1,a^2)=(1-\alpha_t)Q_t^k(s,a^1,a^2)+\alpha_t[r_t^k+\beta\pi^1(s')Q_t^k(s')\pi^2(s')]$更新, 其中 $k=1,2$, $(\pi^1(s'),\pi^2(s'))$是对双矩阵对策$(Q_t^1(s'),Q_t^2(s'))$的混合策略 Nash 平衡点. 序列$\{Q_t^1,Q_t^2\}$收敛于 Nash 平衡点 Q 值(Q_*^1,Q_*^2),证明略.

4 结论

本文提出了在一般和随机对策理论框架下的多智能体学习方法,以 Nash 平衡点作为学习目标. 给出了智能体可以从任意初始 Q 值学习到 Nash 平衡点 Q 值表的多智能体 Q-learning算法,并证明了在一个或多个相同 Nash 点情况下算法的收敛性,为研究和应用多智能体系统提供了理论基础和方法.

参 考 文 献

1 Richard S Sutton, Andrew G Barto. Reinforcement Learning: An Introduction. Cambridge,MA: MIT Press, 1998

2 Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In: Proceedings of the Eleventh International Conference on Machine Learning, New Brunswick, 1994. 157～163

3 Leslie Kaelbling, Michael L Littman, Andrew W Moore. Reinforcement learning: a survey. *Journal of Artificial Intelligence Research*, 1996, (4):237～285

4 Gerhard Weib. Introduction to Distributed Artificial Intelligence. Cambrige, MA: MIT Press, 1998

5 Guillermo Owen. Game Theory, the third edition. San Diego: Academic Press, 1995

6 Jerzy Filar, Koos Vrieze. Competitive Markov Decision Process. Heidelberg, Germany: Springer-Verlag, 1997

欧海涛 上海交通大学博士生. 主要研究领域为多智能体系统和智能控制.

张卫东 上海交通大学自动化系教授,博士生导师. 主要研究领域为鲁棒控制和现场总线技术、多智能体系统.

许晓鸣 上海交通大学自动化系教授,博士生导师,上海交通大学副校长. 研究方向为智能控制.