第 24 卷第 2 期
2007 年 04 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 24 No. 2
Apr. 2007

**文章编号**：1000−8152(2007)02−0317−05

# 一般和博弈中的合作多agent学习

宋梅萍, 顾国昌, 张国印, 刘海波

(哈尔滨工程大学 计算机科学与技术学院, 黑龙江 哈尔滨 150001)

**摘要**: 理性和收敛是多agent学习研究所追求的目标. 在理性合作的多agent系统中提出利用Pareto占优解代替非合作的Nash平衡解进行学习, 使agent更具理性. 另一方面引入社会公约来启动和约束agent的推理, 统一系统中所有agent的决策, 从而保证学习的收敛性. 利用2人栅格游戏对多种算法进行验证, 成功率的比较说明了所提算法具有较好的学习性能.

**关键词**: 多agent学习; 一般和随机博弈; Nash平衡; Pareto占优; Q-学习

**中图分类号**: TP273　　**文献标识码**: A

# Multi-agent learning in cooperative general-sum games

SONG Mei-ping, GU Guo-chang, ZHANG Guo-yin, LIU Hai-bo

(College of Computer Science and Technology, Harbin Engineering University, Harbin Heilongjiang l150001, China)

**Abstract:** Rationality and convergence are two topics in the research on multi-agent learning. A new method called Pareto-Q is proposed with the concept of Pareto optimum, which is more rational than Nash equilibrium with regard to the cooperative system. At the same time, social conventions are also introduced to promise the convergence of learning. When tested on a two-person grid game, the algorithm performs better than the single Q-learning and Nash-Q learning.

**Key words:** multi-agent learning; general-sum game; Nash equilibrium; Pareto optimum; Q-learning

## 1　引言(Introduction)

随着单agent学习研究的成功, 多agent系统的学习问题正成为近年来的研究重点. 由于多agent任务环境下, 个体的回报和环境的转移依赖于所有agent的行为, 从而环境对单个agent而言则更加不确定, 也就很难再描述成一般的Markov过程. 通常的做法是将整个系统看作一个整体, 使环境的转移对系统而言仍然满足Markov特性, 而系统的每个状态又为agent形成一个个的阶段博弈, 从而构成常用来研究多agent学习问题的随机博弈框架.

在该框架下的研究根据采用的博弈解概念可分为用于0和博弈的Minimax-Q学习算法[1], 以及用于一般和博弈的Nash-Q[2,3], FFQ[4], CEQ[5]学习算法等. Minimax-Q的收敛性和有效性已经得到了证明[6], 结果是令人信服的. 而一般和博弈下的各算法则倍受质疑. 如Nash-Q学习算法的条件假设[2,3]过于严格, 收敛不能保证[4]; FFQ学习中合作形式的特殊化, agent无整体理性[5]; CEQ学习的联合动作空间维度灾难[5], 相关均衡解概念[5]的理性程度不高等.

因为一般和博弈是实际应用中最常见的博弈形式, 合作在很多多agent系统中也具有重要的作用, 所以对于合作一般和博弈中的多agent学习算法研究具有实际的意义. 目前, 一般和博弈中应用最广的是Nash-Q学习算法, 但Nash平衡解[3]是非合作情况下的理性最优解概念(如囚徒困境问题), 并不完全适合于合作情况; 另外, Nash-Q学习算法利用两个严格的条件假设[3], 即限定学习过程中的每个对策形势都只有一个全局最优或鞍点Nash解, 来保证所有agent的解选择一致. 这在实际应用中几乎不能实现[4]. 对于博弈中存在多个Nash平衡解的情况, Nash-Q学习算法没有给出具体的解决办法[4].

解选择一致时学习的收敛性可证[3], 所以研究要解决的问题就是理性合作时解概念的确定和解选择的统一. 首先利用Pareto占优解[7] 代替Nash解进行学习以增强agent理性; 其次, 利用合作体内部的社会公约[8]解决解选择问题.

## 2　相关定义(Relative definitions)

在介绍多agent学习的随机博弈框架之前, 需要

收稿日期: 2005−07−11; 收修改稿日期: 2006−05−18.

先给出马尔可夫决策过程(MDP: Markov decision process)和博弈论(GT: game theory)的定义.

### 2.1 MDP定义(Definition of MDP)

很多单agent的强化学习模型可以形式化为马尔可夫决策过程, 其定义如下:

**定义 1**[9] 马尔可夫决策过程即多元组$(S, A, R, T)$, 其中:

1) $S$为环境的离散状态集;

2) $A$为agent的可得离散动作集;

3) $R$为实值奖赏函数: $S \times A \to R$;

4) $T$为状态转移函数: $S \times A \to PD(S)$.

$PD$为状态集$S$上的一个概率分布函数. 函数$T$符合Markov特性, 即对于$a_t$有$P[s_{t+1}|s_t, a_t, \cdots, s_0, a_0] = P[s_{t+1}|s_t, a_t],\ a_t \in A, s_t \in S$.

### 2.2 博弈的定义(Concept of game)

某一状态下的多人交互即形成一个博弈形势. 设局中人集合$N = \{1, 2, \cdots, n\}$, 每个局中人$i$的策略集$S^i$, 其在联合策略下的支付函数$P^i$, $i \in N$. $P^i$定义为诸局中人所取联合策略$(\sigma^1, \sigma^2, \cdots, \sigma^n)$的函数, 其中$\sigma^i \in s^i, i \in N$. 由此定义博弈为

**定义 2**[10] 给定三元组

$$\Gamma =< N, \{S^i\}_{i\in N}, \{P^i\}_{i\in N} > .$$

其中: $N, S^i$均是集合, $i \in N$; 而$P^i$是定义在$S = \prod_{i\in N} S^i$上的实值函数, 则称$\Gamma$为一个博弈.

根据是否容许事先交换、传递信息并订立某种强制性约定, 可将一般和博弈分为非合作博弈和合作博弈[10]. 非合作博弈中常用的解概念是Nash平衡解, 而合作博弈中常用Pareto占优解. 下面给出二者的定义.

**定义 3**[3] Nash平衡是一个$n$元组$\sigma^* = (\sigma_*^1, \cdots, \sigma_*^n)$, 使得对于所有的$\sigma^i \in S^i, i \in N$有

$$P^i(\sigma_*^1, \cdots, \sigma_*^n) \geqslant P^i(\sigma_*^1, \cdots, \sigma_*^{i-1}, \sigma^i, \sigma_*^{i+1}, \cdots, \sigma_*^n).$$

**定义 4**[9] 合作型博弈中, 决策变量$\sigma^* \in S$为Pareto占优的条件是: 不存在其他任何变量$\sigma \in S$, 使对所有的$i = 1, \cdots, n$有$P^i(\sigma) \geqslant P^i(\sigma^*)$, 且至少对一个局中人$j$有$P^j(\sigma) > P^j(\sigma^*)$成立.

### 2.3 多agent随机博弈的定义(Definition of stochastic game)

在上述基础上, 给出多agent随机博弈的形式描述:

**定义 5**[11] 随机博弈即多元组$< n, S, A^{1,\cdots,n}, T, R^{1,\cdots,n} >$, 其中:

1) $n$为agent的数量;

2) $S$为离散状态集;

3) $A^i$为agent $i$的可得动作集, A为所有agent的联合动作空间$A = A^1 \times A^2 \times \cdots \times A^n$;

4) $T$为状态转移函数:$S \times A \times S \to [0, 1]$, 满足$\forall s \in S, \forall \vec{a} \in A, \sum_{s\in S} T(s, \vec{a}, s') = 1$;

5) $R^i$为第$i$个agent的回报函数, $S \times A \to R$. 其中状态转移函数$T$符合Markov特性.

## 3 Pareto-Q算法(Pareto-Q learning algorithm)

### 3.1 Pareto-Q算法形式描述(Formulation description)

单agent Q-学习中Q-函数的调整方式为

$$Q_{t+1}(s, a_t) = (1-\alpha_t)Q_t(s, a_t) + \alpha_t[r_t + \gamma V(s')] \tag{1}$$

其中: $\alpha_t$为随时间$t$衰减的学习速度, $V(s')$为状态值函数:

$$V(s') = \max\{Q_t(s', a)\}. \tag{2}$$

在将其扩展到多agent学习中时, 最直观的方法是将单个agent的动作替换为所有agent的联合动作:

$$Q_{t+1}^i(s, \vec{a}) = (1-\alpha_t)Q_t^i(s, \vec{a}) + \alpha_t[R_t^i(s, \vec{a}) + \gamma V^i(s')]. \tag{3}$$

同时定义新的值函数. 不同算法所采用的值函数定义不同, 如Minimax-Q采用的是矩阵博弈的最大最小值, CEQ采用博弈的相关均衡, 而Nash-Q则是以Nash平衡解来定义值函数.

由此, 定义Pareto-Q学习的更新规则:

$$Q_{t+1}^i(s, a^1, \cdots, a^n) = (1-\alpha_t)Q_t^i(s, a^1, \cdots, a^n) + \alpha_t[r_t^i(s, \vec{a}) + \gamma \mathrm{Pareto}Q_t^i(s')], \tag{4}$$

$$\mathrm{Pareto}Q_t^i(s) = \pi^1(s) \cdots \pi^n(s) \cdot Q_t^i(s). \tag{5}$$

其中: $\pi^1(s) \cdots \pi^n(s)$为Q-值博弈形势$Q_t^1(s), \cdots, Q_t^n(s)$的Pareto占优解.

agent $i$为了在减小通信量的情况下掌握博弈形势, 需要在学习自己Q-函数的同时维护一个关于agent $j (j \neq i)$的Q-函数, 即需要不断更新下式:

$$Q_{t+1}^j(s, a^1, \cdots, a^n) = (1-\alpha_t)Q_t^j(s, a^1, \cdots, a^n) + \alpha_t[r_t^j(s, \vec{a}) + \gamma \mathrm{Pareto}Q_t^j(s')]. \tag{6}$$

### 3.2 制定公约(Creating lexicographic convention)

博弈中通常有多个Pareto占优解, 且不同agent对各占优解的偏好不同, 所以如何在多个合作agent间统一占优解的选择, 从而保证学习的收敛效果就是需要进一步考虑的问题. 通常的解决办法有两种[12]: 一种是通信协商, 如引入可信第3方[13]来促

使达到博弈双方一致认可的结果. 但是这一方法显然不太适用于多agent学习过程. 因为一方面过多的通信会减慢agent的决策从而大大影响学习速度; 另一方面, 同一状态下不同时刻的协商结果可能不同, 也会影响学习的收敛效果. 另一种方法是制定社会公约, 在agent之间形成一个公共知识[8], 从而对各agent的动作选择形成一定的约束.

Jelle曾为多人合作博弈给出一个简单的公约[12]. 假设agent有能力互相确认, 则可以利用下面几条假设来设定一个简单的规则:

· agent集有序;

· 每个agent的动作集有序;

· 这种顺序在agent间是公共知识.

根据该公约, 选择一个最优联合动作的过程如下: 第1个agent选择其动作顺序中出现的第1个Nash平衡解对应的动作; 第2个agent则在给定agent 1的选择后, 在自己的动作顺序中选第1个最优动作; 依次向后, 直到所有的agent选择完毕.

下面针对常用到的几种一般博弈形势对公约进行修正(见图1).

| Game 1 | Left | Right |
|---|---|---|
| Up | 10, 9 | 0, 3 |
| Down | 3, 0 | −1, 2 |

（a）

| Game 2 | Left | Right |
|---|---|---|
| Up | 5, 5 | 0, 6 |
| Down | 6, 0 | 2, 2 |

（b）

| Game 3 | Left | Right |
|---|---|---|
| Up | 10, 9 | 0, 3 |
| Down | 3, 0 | 2, 2 |

（c）

图 1　3种常见一般和对策

Fig. 1　Three typical games of general-sum

对于双矩阵博弈1, 只存在一个最优Nash平衡解(10, 9), 平衡解的选择不冲突.

对于双矩阵博弈2(囚徒困境问题), 只存在一个鞍点解(2,2), 但存在占优于鞍点的Pareto占优解(5,5), 此时, 定义Pareto占优解≥鞍点解.

对于双矩阵博弈3, 同时存在最优解(10,9)和鞍点解(2,2), 定义最优解≥ 鞍点解.

由此, 给出新订立的社会公约为:

· 各种博弈解有序(最优Nash解≥Pareto占优解≥鞍点Nash解);

· agent集有序;

· 每个agent的动作集有序;

· 这种顺序在agent间是公共知识.

公约中的第1条规则可以实现解概念的确定, 确保所有agent在博弈的Pareto占优解集中进行选择; 第2,3条与原有规则功能相同, 可以保证agent在给定解集中的解选择一致.

### 3.3　算法描述(Algorithm description)

各agent的学习过程可以描述为下面的步骤:

1) 初始化: 对所有的$s_0 \in S, a^i \in A^i, i=1,2,\cdots,n$,令$Q_0^i(s_0,a^1,\cdots,a^n)=0, Q_0^j(s_0,a^1,\cdots,a^n)=0, j\neq i$;

2) 重复下面的操作:

a) 观察当前状态$s_t$;

b) 根据社会规则和自身保留的$Q_t^i(s_t,a^1,\cdots,a^n)$和$Q_t^j(s_t,a^1,\cdots,a^n),(j\neq i)$选择Pareto占优动作;

c) 观察新状态, 各agent彼此的回报及前一时刻所采取的动作;

d) 根据式(4)和式(6)来更新自己和其他agent的Q-函数值.

直到满足结束条件.

## 4　实验及结果分析(Experiment and results)

图3是agent学习中的常用栅格游戏实验, 它具备动态环境的所有关键因素, 即基于位置或状态的动作, 定性的转移(agent的移动)以及即时和长期回报. 游戏的任务是各agent在不冲突的前提下, 以最少的移动步数到达各自的目标点. 从图2的状态出发, 先到达目标点的agent可获得一定的回报. 若两者同时到达, 则都可以得到与前面等量的回报, 即一个agent的成功不会阻止另一个agent的成功, 二者非对抗关系, 这所形成的就是一个一般和博弈. 假设agent可以观察彼此的前一动作, 当前状态(两个agent的联合位置)以及彼此的回报, 且初始状态下, 各agent不知道自己的目标点位置.

本文采用VC++在Windows 2000平台上实现了该栅格游戏, 并采用它对一般和博弈中多agent学习的不同算法进行了对比和分析.

### 4.1　随机博弈表达(Representation as stochastic games)

agent $i$的动作空间为

$$A^i=\{\text{left},\text{right},\text{up},\text{down}\}(i=1,2);$$

状态空间为

$$S=\{(0,1),(0,2),\cdots,(8,7)\}.$$

其中: 状态$s=(l^1,l^2)$表示两个agent的联合位置, agent $i$的位置用图3中的位置号表示.

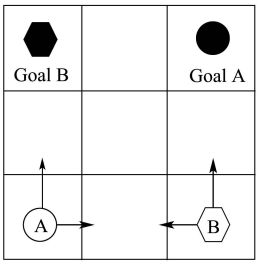


图 2 删格游戏示意图

Fig. 2 Grid game

| | | |
|---|---|---|
| 6 | 7 | 8 |
| 3 | 4 | 5 |
| 0 | 1 | 2 |

图 3 位置号示意图

Fig. 3 Positions of agent

强化信号的设定为: agent到达目标时, 强化信号为100; 与其他agent相撞时, 强化信号为–1, 且两个agent都回到前一位置; 否则, 强化信号为0. 即

$$r_t^i = \begin{cases} 100, & L(l_t^i, a_t^i) = \mathrm{Goal}_i, \\ -1, & L(l_t^1, a_t^1) = L(l_t^2, a_t^2), \\ 0, & \text{其他}. \end{cases} \tag{7}$$

学习速度定义为访问次数的倒数, 即$\alpha_t(s, a^1, a^2) = \dfrac{1}{n_t(s, a^1, a^2)}$, $n_t(s, a^1, a^2)$为$(s, a^1, a^2)$被访问的次数.

agent学习过程中进行动作选择时, 采用$\varepsilon-$贪婪策略, 以$\varepsilon_t(s) = \dfrac{1}{1 + n_t(s)}$选择探索动作, 以$(1 - \varepsilon_t(s))$选择Pareto占优解动作.

### 4.2 实验结果与性能比较(Results and performance analysis)

设定学习结束条件为学习次数到达40000次, 因为该游戏中的状态-动作对总数为424, 所以每个状态-动作对平均被访问的次数为95次, 当结束时, 学习速度为$\alpha_t = \dfrac{1}{95} \approx 0.01$, 新访问的结果就不太能够影响已学得的Q-函数值. 利用3种学习agent进行实验, 即Single agent,First Nash agent和Pareto agent. 其中, Single agent采用的是单agent学习算法, 即Q-学习; First Nash agent采用的是Nash-Q学习算法, 并人为规定系统中所有的agent选择第1个Nash平衡解策略; Pareto agent采用的是本文所提出的Pareto-Q学习算法, 并利用社会公约统一系统的解选择.

实验结果如表1所示, 每种情况都运行100次, 计算实现最优联合路径的成功率.

表 1 学习结果比较

Table 1 Comparison of learning results

| 学习策略 | | 学习结果 |
|---|---|---|
| Agent1 | Agent2 | 实现最优联合路径的成功率 |
| Single | Single | 23% |
| First Nash | Single | 45% |
| Pareto | Single | 54% |
| First Nash | First Nash | 100% |
| Pareto | Pareto | 100% |

由结果可以看出, 两个agent都采用单agent Q-学习时, 实现成功率为23%, 这是单agent不对其他agent建模的原因所致; Nash-Q agent与单agent时, 实现成功率为45%, 这与Hu Junling的实验结果比较接近, 较第一种情况有了很大的提高; 而相同条件设置下的Pareto-Q agent与单agent时, 实现成功率为54%, 比Nash-Q又有了一定的提高; 当两个agent的类型相同, 且选解一致时, 后两种agent的实现成功率都可以达到100%. 但是在Nash-Q学习算法的实现过程中, Hu Junling是利用一个外界的告知来规定agent选择第1个或第2个Nash平衡解, 对于非合作情况下这个外界的告知的来源未作说明.

## 5 结论及后期工作(Conclusions and the future works)

Pareto-Q学习算法所使用的Pareto占优解概念与Nash平衡解相比更适用于合作的一般和博弈, 使合作的agent更具理性. 另外, 利用社会公约的方法统一各agent的决策, 取消了Nash-Q学习算法中的条件假设. 使算法对学习过程中的对策形式没有特别要求, 更便于应用. 社会公约使所有agent的解选择达到一致, 保证了学习的收敛性. 在Hu Junling对解选择一致时学习收敛性做出的理论证明基础上, 文章中的实验结果也进一步提供了有利的依据.

与FFQ和CEQ相比较, 学习目标的选择使得Pareto-Q中的agent在个体和整体利益方面都显得更为理性.

但是, 目前关于这方面的工作还在进行中, 仍有几个问题需要解决:

第一, 合作中的背叛. 这里没有考虑合作agent间的背叛行为, 对于这一问题, 可以利用惩罚机制(n tit-for-a-tat)做深入的研究.

第二, 信度分配问题. 因为不同理性agent对不同的Pareto占优解的倾向程度不同, 在实现统一的选择后, 如何平衡agent间的利益分配也是需要进一步考虑的问题.

第三, 不存在纯策略Nash平衡时的解选择问题. 该问题在其他算法中也没有进行讨论, 仍需进一

步研究.

第四, 复杂任务的状态, 动作空间问题. 在应用到复杂任务中时, 状态, 动作空间的组合爆炸及此时对策解的计算需继续研究.

## 参考文献(References):


[1] LITTMAN M L. Markov games as a framework for multi-agent reinforcement learning[C] // *Proc of the Eleventh Int Conf on Machine Learning*. New Brunswick, NJ San Mateo, CA: Morgan Kaufmann Publishers, 1994: 157 – 163.

[2] HU Junling, LITTMAN M L. Multiagent reinforcement learning: theoretical framework and an algorithm[C] // *Proc of the Fifteenth Int Conf on Machine Learning*. Madison, Wisconson, San Mateo, CA: Morgan Kaufmann Publishers, 1998: 242 – 250.

[3] HU Junling, WELLMAN M P. Nash Q-Learning for general-sum stochastic games[J]. *J of Machine Learning Research*, 2003, 4(6): 1039 – 1069.

[4] LITTMAN M L. Friend or foe Q-learning in general-sum markov games[C] // *Proc of the Int Conf on Machine Learning*. Williams Colledge, MA, San Mateo, CA: Morgan Kaufmann Publishers, 2001(a): 322 – 328.

[5] GREENWALD A, HALL K. Correlated Q-learning[C] // *Proc of the Twentieth Int Conf on Machine Learning*. Washington DC, USA: AAAI Press, 2003: 242 – 249.

[6] LITTMAN M L, SZEPESVARI C. A generalized reinforcement learning model: Convergence and applications[C] // *Proc of the 13th Int Conf on Machine Learning*. Bari, Italy, San Mateo, CA: Morgan Kaufmann Publishers, 1996: 310 – 318.

[7] DEB K. *Multi-objective evolutionary algorithms: Introducing bias among pareto-optimal solutions,KanGAL report 99002*[R]. Kanpur, India: Indian Institute of Technology, 1999.

[8] GEANAKOPLOS J. Common knowledge[J]. *J of Economic Perspectives*, 1992, 6(4): 53 – 82.

[9] 高阳, 周志华, 何佳洲, 等. 基于Markov对策的多Agent强化学习模型及算法研究[J]. 计算机研究与发展, 2000, 37(3): 257 – 263.
(GAO Yang, ZHOU Zhihua, HE Jiazhou, et al. Research on Markov game-based multiagent reinforcement learning model and algorithms[J]. *J of Computer Research and Development II*, 2000, 37(3): 257 – 263.)

[10] 刘德铭, 黄振高. 对策论及其应用[M]. 长沙: 国防科技大学出版社, 1994.
(LIU Deming, HUANG Zhengao. *Game Theory and Its Applications*[M]. Changsha: National University of Defense Technology Press, 1994.)

[11] BOWLING M,VELOSO M. Existence of multiagent equilibria with limited agents[J]. *J of Artificial Intelligence Research*, 2004, 22(2): 353 – 384.

[12] KOK J R, SPAAN M T J,VLASSIS N. An approach to noncommunicative multiagent coordination in continuous domains[C] // WIERING M. *Proc of the Twelfth Belgian-Dutch Conf on Machine Learning*. Utrecht, Netherlands: University of Utrecht, 2002: 46 – 52.

[13] 张虹, 邱玉辉. 一个基于对策论的协商模型[J]. 南京大学学报(自然科学), 2001, 37(2): 159 – 164.
(ZHANG Hong, QIU Yuhui. A negotiation model that based game theory[J]. *J of Nanjing University(Natural Science)*, 2001, 37(2): 159 – 164.)


## 作者简介:


**宋梅萍**　(1978—), 女, 哈尔滨工程大学计算机科学与技术学院博士研究生, 主要研究领域为多机器人系统、体系结构等, E-mail: smping@163.com ;

**顾国昌**　(1946—), 男, 哈尔滨工程大学计算机科学与技术学院教授, 博士生导师, 研究领域为机器人系统、体系结构及智能规划与决策等, E-mail: guguochang@hrbeu.edu.cn;

**张国印**　(1962—), 男, 哈尔滨工程大学计算机科学与技术学院教授, 博士生导师, 研究领域为主要研究领域为智能控制、智能机器人、嵌入式系统、网络安全、电子商务安全、智能搜索引擎等, E-mail: zhangguoyin@hrbeu.edu.cn;

**刘海波**　(1976—), 男, 哈尔滨工程大学计算机科学与技术学院讲师, 博士, 研究领域为多智能体系统、机器人智能, E-mail: liuhaibo@hrbeu.edu.cn .