计算机科学2004Vol. 31№. 4

多代理最优响应Q学习及收敛性证明

张化祥　黄上腾
（上海交通大学计算机科学与工程系　上海200030）

摘　要　在分析了多代理强化学习的基础上，提出了一种基于对手策略假设的代理最优响应强化学习规则，并证明了当对手策略满足一定条件时，基于该学习规则的Q值收敛。实验结果与理论证明相一致。

关键词　强化学习，Markov对策，收敛

Multiagent Optimal Response Q-learning and its Convergence

ZHANG Hua-Xiang　HUANG Shang-Teng
(Department of Computer Science & Engineering, Shanghai Jiaotong Univ., Shanghai 200030)

Abstract　Based on analysis of multiagent reinforcement learning, an agent optimal response learning rule is proposed provided the assumptions of opponents' policy. Q values have been proved to be convergent if opponents' policy satisfies certain restrictions, and experimental results of grid games are consistent with the convergence proof.

Keywords　Reinforcement learning, Markov games, Convergence

1　引言

强化学习[1]近年来被越来越多地应用到多代理领域的研究。代理间的交互导致问题的复杂，同时多代理环境不是静态系统，因此将强化学习技术应用到多代理环境，应对代理的学习规则进行修改。基于Q学习[3]提出的多代理强化学习规则主要有Littman的minimax-Q学习[4]、Hu与Wellman的Nash Q学习[5,6]、Claus与Boutilier[7]的全合作多代理Q学习、Bowling与Veloso的可变学习率的Q学习[11]、Littman的friend-or-foe Q(FF-Q)学习[12]及Greenwald的关联Q学习(CE-Q)[8]。

两代理零和对策下，Littman的minimax-Q学习基于悲观假设，代理学习最差情况下的最优行动策略。minimax-Q学习满足收敛性[2]。Hu与Wellman针对一般和对策提出Nash-Q学习。代理的行动策略为选择对策每一个阶段的Nash均衡点。满足一定约束条件时Nash Q学习收敛。但有多个Nash均衡点存在时，很难保证代理选择相同的Nash均衡点且选择最优的均衡点。Littman[12]将一般和对策中的代理分成两种："朋友"与"敌人"。"朋友"的策略是使代理的收益最大化，"敌人"的策略是使代理的收益最小化。基于两种极端假设的FF-Q算法能保证收敛。Greenwald提出的基于关联Nash均衡点的CE-Q，基本思想类同于Nash-Q，只是代理在Q学习时选择关联均衡处代理的单阶段对策期望收益。与Nash-Q比较，CE-Q计算方便，CE-Q均衡点的计算只需要线性规划。CE-Q依然存在均衡点选择的问题。只有一个均衡点存在时，基于对手选择均衡点策略的假设，CE-Q收敛。若对手采取不同的学习策略，比如Nash-Q、Q学习或FF-Q时，能否保证Q值的收敛？

以上多代理学习需要联合行动信息，同时需要存储每个阶段对策中所有代理的Q值。这样做需要大量的存储空间并增加代理间的通讯。Suematsuhe和Hayashi[10]提出了一种扩展式最优响应学习，减少了Q值的存储，但没有给出形式化描述，验证代理的策略或Q值函数的收敛性。

本文提出一种基于代理最优响应的多代理强化学习规则。对手策略满足一定假设条件时，基于该策略学习的代理，Q值能保证收敛。

2　多代理最优响应策略强化学习

对策论中增加状态空间后，代理的每一步决策都导致状态的转移，此时的对策就是一个随机对策。具有Markov特性的随机对策是称为Markov对策(MG)。MG是研究具有离散时间特性多代理协同的重要理论框架。

MG是一个五元组$(I,S,\{A_i\}_{i\in I},P,\{R_i\}_{i\in I})$。$I$是对策代理集合；$S$是状态集合；$A_i$是第$i$个代理可选行动的集合(在状态$s\in S$，第$i$个代理可选行动的集合记为$A_i(s)$)。对策中代理在状态$s$的联合行动集合记为$A(s)=A_1(s)\times A_2(s)\times\cdots\times A_n(s)$($n$为代理数)，有联合行动向量$\vec{a}(s)\in A(s)$；$P$为状态转移函数；$R_i(s,\vec{a})$表示第$i$个代理在状态$s$采取联合行动$\vec{a}$的立即收益。一般将$R_i(s,\vec{a})$记为$R_i(s,a_i,a_{-i})$，其中$a_i\in A_i$表示第个代理的行动，$a_{-i}$表示除第$i$个代理的其它代理的联合行动。

混合策略意义下的MG是指代理i的混合策略$\pi_i:A_i\to\Pi_i(A_i)$，表示代理i在其行动集合A_i上的一种概率分布。代理的混合策略组合$(\pi_1,\pi_2,\cdots,\pi_n)$构成系统的一个混合策略。不同状态下有不同的混合策略。在状态s采用混合策略$(\pi_1(s),\pi_2(s),\cdots,\pi_n(s))$时代理$i$的立即期望收益为

$$Er_i(s,\pi_1,\cdots,\pi_n)=\sum_{a_i\in A_i}\sum_{a_{-i}\in A/A_i}\pi_1\cdots\pi_n R_i(s)。$$

带折扣MG的目标是最大化代理的累计折扣收益。设有折扣因子γ，代理的策略组合为$(\pi_1,\pi_2,\cdots,\pi_i,\cdots,\pi_n)$，给定初始状态$s$，则代理$i$的收益

$$V_i(s,\pi_1,\cdots,\pi_n)=\sum_{t=0}^{\infty}\gamma^t Er_i(\pi_1,\cdots,\pi_n,s_0=s)$$

代理i的目标就是最大化$V_i(s,\pi_1,\cdots,\pi_n)$。

MG中，代理i的最优响应是指对于代理I/i(i的补集)

的联合行动策略 π_{-i}，代理 i 所有最优策略的集合。记为 $BR_i(\pi_{-i})$

$$BR_i(\pi_{-i})=\{\pi_i\in\Pi_i\mid V_i(\pi_i,\pi_{-i})\geqslant V_i(\pi_i',\pi_{-i}),\forall\ \pi_i'(\neq\pi_i)\in\Pi_i\}$$

代理每一步都采取最优的行动策略，此时的 Q 学习称为最优响应 Q 学习。

给定对手策略假设 $\pi_{-i}(s)$，代理 i 有最优策略 $\pi_i^*(s)$，使得

$\pi_i^*(s)=\arg\max_{\pi_i(s)\in\Pi_i(s)}V_i(s,\pi_i,\pi_{-1})$。此时，

$$Q_i^*(s,\vec{a})=r_i(s,\vec{a})+\gamma\sum_{s'\in S}P_{ss'}(\vec{a})V_i)(s',\pi_i^*,\pi_{-i})),$$

针对迭代的每一步，代理 i 应选择当前迭代的最优策略。此时 Q 学习更新规则为

$$Q_i^{t+1}(s;\vec{a})=(1-\alpha)Q_i^t(s,\vec{a})+\alpha(r_i(s,\vec{a})+\gamma\max_{\pi_i(s')\in\Pi_i(s')}\pi_i(s')Q_i^t(s',\vec{a})\pi_{-i}(s'))$$

$\pi_{-i}(s)$满足一定的假设条件且 $t\to\infty$ 时，$Q_i^t(s,\vec{a})$收敛到 $Q^*(s,\vec{a})$。

3 收敛性定理

引理1[9] 迭代 $(*)Q_{t+1}(x)=(1-\alpha_t(x))Q_t(x)+\alpha_t(x)[P_tQ_t](x)$。假设 $Q_{t+1}(x)=(1-\alpha_t(x))Q_t(x)+\alpha_t(x)[P_tQ^*](x)$产生的$\{Q_t(x)\}$序列以概率1收敛到 Q^*。其中 P_t 为映射 $P_t:\mathbb{Q}\to\mathbb{Q}$。如果下面的条件满足：$0<\gamma<1$和序列$\{\lambda_t\mid\lambda_t\geqslant0\}$以概率1收敛到0。若$\|P_tQ-P_tQ^*\|\leqslant\gamma\|Q-Q^*\|+\lambda_t$ 对$\forall\ Q\in\mathbb{Q}$成立，且 $\alpha_t(x)$满足$0\leqslant\alpha_t(x)<1$，$\sum_{t=0}^{\infty}\alpha_t(x)=\infty$，$\sum_{t=0}^{\infty}\alpha_t^2(x)<\infty$，则迭代$(*)$产生的序列$\{Q^t(x)\}$当 $t\to\infty$时，以概率1收敛到 $Q^*(x)$。

定理1 n 个代理的 MG，代理 i 假定对手采用固定的联合策略 π_{-i}，且该假设正确。基于该假定，则 Q 学习更新规则

$$Q_i^{t+1}(s,\vec{a})=(1-\alpha)Q_i^t(s,\vec{a})+\alpha(r_i(s,\vec{a})+\gamma\max_{\pi_i(s')\in\Pi_i(s')}\pi_i(s')Q_i^t(s',\vec{a})\pi_{-i}(s'))$$

产生的序列$\{Q_i^t(s,\vec{a})\}$收敛到 $Q_i^*(s,\vec{a})$对$\forall\ s\in S,\forall\ \vec{a}\in A$ 成立。其中

$$Q_i^*(s,\vec{a})=r_i(s,\vec{a})+\gamma\sum_{s'\in S}P_{ss'}(\vec{a})V_i(s',\pi_i^*,\pi_{-i}))$$

证明：定义 $P_tQ_i^t(s,\vec{a})=r_i(s,\vec{a})+\gamma\max_{\pi_i(s')\in\Pi_i(s')}\pi_i(s')Q_i^t(s',\vec{a}')\pi_{-i}(s')$。有$\|P_tQ-P_tQ^*\|\leqslant\max_{s\in S}|P_tQ_i^t(s,\vec{a})-P_tQ_i^*(s,\vec{a})|$。其中 P_t 是空间 Q_t 到 Q_t 的映射。

同理有 $P_tQ_i^*(s,\vec{a})=r_i(s,\vec{a})+\gamma\max_{\pi_i\in\Pi_i}\pi_i(s')Q_i^*(s',a^{\rightarrow})\pi_{-i}(s')$。

$$|P_tQ_i^t(s,\vec{a})-P_tQ_i^*(s,\vec{a})|=\gamma|\max_{\pi_i(s')\in\Pi_i(s')}\pi_i(s')Q_i^t(s',\vec{a}')\pi_{-i}(s')-\max_{\pi_i(s')\in\Pi_i(s')}\pi_i(s')Q_i^*(s',\vec{a}')\pi_{-i}(s')|$$

$$\leqslant\gamma|Q_i^t(s',\vec{a})-Q_i^*(s',\vec{a})|$$

已经有 $Q_i^*(s,\vec{a})=r_i(s,\vec{a})+\gamma\sum_{s'\in S}P_{ss'}(\vec{a})V_i(s',\pi_i^*,\pi_{-i})$

$=r_i(s,\vec{a})+\gamma E_{s'}(V_i(s',\pi_i^*,\pi_{-i}))$

$$E[P_tQ_i^*](s,\vec{a})=E(r_i(s,\vec{a})+\gamma\max_{\pi_i\in\Pi_i,\pi_i}(s')Q_i^*(s',\vec{a}')\pi_{-i}(s'))=r_i(s,\vec{a})+\gamma E(\max_{\pi_i\in\Pi_i}\pi_i(s')Q_i^t(s',\vec{a}')\pi_{-i}(s'))$$

因为 $V_i(s',\pi_i^*,\pi_{-i})=\max_{\pi_i\in\Pi_i}\pi_i(s')Q_i^*(s',\vec{a}'\to)\pi_{-i}(s')$

故 $Q_i^*=E[P_tQ_i^*]$。引理1的两个条件都满足，所以说序列$\{Q_i^t(s,\vec{a})\}$收敛到 $Q_i^*(s,\vec{a})$对$\forall\ s\in S,\forall\ \vec{a}\in A$ 成立。

定理2 n 个代理的 MG，代理 i 假定对手采用的联合策略为 $\pi_{-i}'(s)$，并且对$\forall\ s\in S$，当 $t\to\infty$时，$\pi_{-i}'(s)$收敛到 $\pi_{-i}^*(s)$（其中 t 为迭代次数）。基于该假定且该假定正确的情况下，Q 学习更新规则

$$Q_i^{t+1}(s,\vec{a})=(1-\alpha)Q_i^t(s,\vec{a})+\alpha(r_i(s,\vec{a})+\gamma\max_{\pi_i(s')\in\Pi_i(s')}\pi_i'(s')Q_i^t(s',\vec{a})\pi_{-i}'(s'))$$

产生的序列$\{Q_i^t(s,\vec{a})\}$收敛到 $Q_i^*(s,\vec{a})$对于$\forall\ s\in S,\forall\ \vec{a}\in A$ 成立。其中

$$Q_i^*(s,\vec{a})=r_i(s,\vec{a})+\gamma\sum_{s'\in S}P_{ss'}(\vec{a})V_i(s',\pi_i^*,\pi_{-i}^*))$$

证明：定义 $P_tQ_i^t(s,\vec{a})=r_i(s,a\to)+\gamma\max_{\pi_i\in\Pi_i}\pi_i(s')Q_i^t(s',\vec{a}')\pi_{-i}(s')$。其中 P_t 是空间 Q_t 到 Q_t 的映射。有 $P_tQ_i^*(s,\vec{a})=r_i(s,\vec{a})+\gamma\max_{\pi_i\in\Pi_i}\pi_i(s')Q_i^*(s',\vec{a}')\pi_{-i}^*(s')$。

$$|P_tQ_i^t(s,\vec{a})-P_tQ_i^*(s,\vec{a})|=\gamma|\max_{\pi_i\in\Pi_i}\pi_i(s')Q_i^t(s',\vec{a}')\pi_{-i}(s')-\max_{\pi_i\in\Pi_i}\pi_i(s')Q_i^*(s',\vec{a}')\pi_{-i}^*(s')|\leqslant\gamma\max_{\pi_i\in\Pi_i}|\pi_i(s')Q_i^t(s',\vec{a}')\pi_{-i}(s')-\pi_i(s')Q_i^*(s',\vec{a}')\pi_{-i}^*(s')|$$

因为 $|\pi_i(s')Q_i^t(s',\vec{a}')\pi_{-i}(s')-\pi_i(s')Q_i^*(s',\vec{a}')\pi_{-i}^*(s')|=$

$|\pi_i(s')Q_i^t(s',\vec{a}')\pi_{-i}(s')-\pi_i(s')Q_i^*(s',\vec{a}')\pi_{-i}(s')+\pi_i(s')Q_i^*(s',\vec{a}')\pi_{-i}(s')-\pi_i(s')Q_i^*(s',\vec{a}')\pi_{-i}^*(s')|$

$\leqslant|\pi_i(s')Q_i^t(s',\vec{a}')\pi_{-i}(s')-\pi_i(s')Q_i^*(s',\vec{a}')\pi_{-i}(s')|+|\pi_i(s')Q_i^*(s',\vec{a}')\pi_{-i}(s')-\pi_i(s')Q_i^*(s',\vec{a}')\pi_{-i}^*(s')|$

$=|\pi_i(s')(Q_i^t(s',\vec{a}')-Q_i^*(s',\vec{a}'))\pi_{-i}(s')|+|\pi_i(s')Q_i^*(s',\vec{a}')(\pi_{-i}(s')-\pi_{-i}^*(s'))|$

于是 $|P_tQ_i^t(s,\vec{a})-P_tQ_i^*(s,\vec{a})|\leqslant\gamma\max_{\pi_i\in\Pi_i}|\pi_i(s')(Q_i^t(s',\vec{a}')-Q_i^*(s',\vec{a}'))\pi_{-i}(s')|+\gamma\max_{\pi_i\in\Pi_i}|\pi_i(s')Q_i^*(s',\vec{a}')(\pi_{-i}(s')-\pi_{-i}^*(s'))|$

$\leqslant\gamma|Q_i^t(s',\vec{a})-Q_i^*(s',\vec{a})|+\delta_t$

很明显 $t\to\infty,\delta_t\to0(\delta_t=\gamma\max_{\pi_i\in\Pi_i}|\pi_i(s')Q_i^*(s',\vec{a}')(\pi_{-i}(s')-\pi_{-i}^*(s'))|)$

同定理1，可以证明 $Q_i^*=E[P_tQ_i^*]$

于是$\forall\ s\in S,\forall\ \vec{a}\in A$，$\{Q_i^t(s,\vec{a})\}$收敛的两个条件满足。所以有$\forall\ s\in S,\forall\ \vec{a}\in A$，$\{Q_i^t(s,\vec{a})\}$收敛到 $Q_i^*(s,\vec{a})$。

基于定理1和2的结论，代理在线学习时，需要建立对手代理的初始策略模型，并在学习的过程中不断地逼近对手的实际策略。

学习过程中，代理不需要计算 Nash 均衡点或关联均衡点。降低了计算复杂性；同时代理不需知道对手的 Q 值函数，减少了存储。另外在学习中不存在选择均衡点的问题。

4 仿真试验

试验环境如图1所示。代理 A1、A2的目标分别为 G1和 G2。每个代理有四种行动（上、右、下、左），并且每次只能移动一个格子。若移向边界，代理将会停留在原来的格子不动，得

－1分；若两代理移向相同的格子，则都停留在原来的格子不动，同时得－1分。其它情况，代理从一个格子移到另一格子的立即收益为0。假设两代理同时选择行动，代理的立即收益及行动可观察。当代理达到目标时，立即收益为100分，另一代理的收益为0；若同时到达目标，两代理各得100分。代理通过学习，找到到达目标的最短路径。

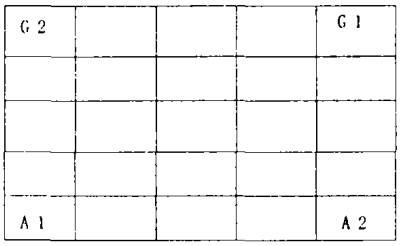

图1　网格对策

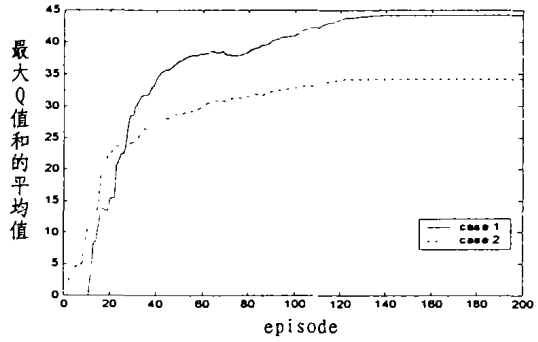

图2　最大Q值和的平均值

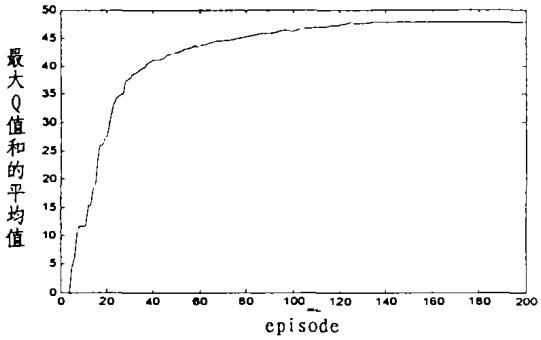

图3　A2策略收敛时A1的Q值收敛

代理从各自的起点出发。当有一个代理到达目标点时，称为一个 episode 结束。然后随机地将代理放到非其目标点的格子里，开始新的 episode。

所有代理采用相同的学习更新规则学习最优策略，且满足定理2收敛性条件的多代理Q学习，如 minimax-Q、Nash-Q、CE-Q、FF-Q 都证明了收敛性，与定理2的结论一致。

图2表明当 A2采用固定的策略时 A1代理 Q 值的收敛性。设置 A2的固定策略为：case 1：除 G2外，A2 在每一个格子选择每个方向的概率都为0.25。case 2：除 G2外，A2 在每一个格子选择向上移动的概率为2/3，选择其它三个方向的概率各为1/9。两种情况下都验证了 A1 Q 值的收敛性。最大Q值和的平均值为 $\sum_{i=1}^{24}\max_a Q(s,a)/24$（24为 A1的状态数）。多次试验结果表明给定 A2的固定策略，A1的 Q 值总是收敛的。

假定 A2的行动策略为向上、右、下、左移动的概率分别为：$\frac{1}{4}+(\frac{1}{5})^t$、$\frac{1}{4}-(\frac{1}{5})^t$、$\frac{1}{4}+(\frac{1}{6})^t$、$\frac{1}{4}-(\frac{1}{6})^t$，其中 t 为 episode 数。A2的行动策略满足定理2的条件，即当 $t\to\infty$ 时其策略收敛。图3的结果表明 A1的 Q 值收敛。

我们比较了对手采用不同的学习算法的性能。结果表明基于不同假设，代理采用不同学习策略的性能各不相同。基于正确假设的情况下代理学习将会取得最好的性能。

结束语　多代理系统中，基于对手不同的策略假设，代理采用最优响应 Q 学习更新规则时，当对手策略满足一定条件，且关于对手策略的假设正确时，代理学习的 Q 值能保证收敛。同时，只有对手代理采用与代理本身相同的学习更新规则时，代理才能取得最好的学习性能。

参 考 文 献

1 Sutton R S, Barto A. Reinforcement Learning: An Introduction [M]. MIT Press, Cambridge, MA

2 Singh S, Jaakkola T, Littman M L. Szpesvari C. Convergence results for single-step on policy reinforcement-learning algorithms. Machine Learning Journal, 2000, 38(3): 287～308

3 Watkins C J C H. Learning from Delayed Rewards: [Ph. D. thesis]. Cambridge, UK: Cambridge University, 1989

4 Littman M L. Markov games as a framework for multi-agent reinforcement learning. In: 11th ICML, New Brunswick, 1994. 157～163

5 Hu J, Wellman M P. Multiagent reinforcement learning: Theoretical framework and an algorithm. In: 15th ICML, p242～250

6 Hu J, Wellman M P. Nash Q-Learning for General-Sum Stochastic Games. Journal of Machine Learning research, 2003, (1): 1～30

7 Claus C, Boutilier C. The dynamics of reinforcement learning in cooperative multiagent systems In: Proc. of the Fifteenth National Conf. on Artificial Intelligence, 1998

8 Greenwald A, Hall K, Serrano R. Correlated-Q learning. In: NIPS Workshop on Multiagent Learning, 2002Craig Boutilier Sequential optimality and coordination in multiagent systems. In: 16th Intl. Joint Conf. on Artificial Intelligence, Stockholm, 1999. 478～485

9 Szepesvari C, Littman M L. A Unified Analysis of Value-Function-Based reinforcement Learning algorithms [J]. Neural computation, 1999, 11(8): 2017～2060

10 Suematsu N, Hayashi A. A Multiagent Reinforcement Learning Algorithm using Extended Optimal Response. In: Proc. of the First Intl. Joint Conf. on Autonomous Agents & Multiagent Systems, Bologna, Italy, 2002. 370～377

11 Bowling M, Veloso M. Multiagent learning using a variable learning rate. Artificial Intelligence, 2002, 136: 215～250

12 Littman M L. Friend-or-foe Q-learning in general-sum games. In: 18th ICML, Williams college, MA, 2001. 322～328