文章编号: 1001-0920(2002)01-0029-04

基于独立学习的多智能体协作决策

李晓萌, 杨煜普, 许晓鸣
(上海交通大学 自动化研究所, 上海 200030)

摘 要: 联合学习模式是实现多智能体协作决策的有效方法, 但是当智能体信息不完备时, 这一方法难以适用。为此, 在智能体独立学习的基础上提出一种多智能体协作决策方法。以网格对策为例, 仿真证明了这一方法的有效性。

关键词: 多智能体强化学习; 独立学习; Markov 协作决策过程
中图分类号: TP 18 文献标识码: A

Multiagent cooperative decision making based on independent learning

LI Xiao-meng, YANG Yu-pu, XU Xiao-ming
(Institute of Automation, Shanghai Jiaotong University, Shanghai 200030, China)

Abstract: Although joint learning is an efficient method to implement multiagent cooperative decision, it is unsuccessful when agent has imperfect information. The method of agents' independent learning which acts as the base of multiagent cooperative decision is put forward. The experiment of grid games shows the efficiency.

Key words: multiagant reinforcement learnning; independent learning; Markov cooprative decision process

1 引 言

关于多智能体强化学习的研究, 近年来引起人们广泛的兴趣。Littman 基于零和对策提出了达到平衡点的学习算法[1]。Hu 和 Wellman 给出了在非零和对策基础上的算法, 并证明了这种算法的收敛性[2]。上述两种算法的共同特点是智能体采用联合行动, 且都具有彼此间的完备信息(对策结构、状态转移概率、奖赏函数)。Claus 和 Boutilier 研究了协作的多智能体决策过程[3], 比较了智能体独立学习和联合学习的差别, 但无法保证这两种学习收敛到平衡点, 对此他们给出一些加强联合学习的建议。另外, Wolpert 等为每个智能体建立了各自的奖赏函数[4], 但该方法使学习过程的分析复杂化, 并增加了计算代价。基于智能体强化学习的应用包括以网络节点为智能体的网络路由方法[5]、电梯群控[6]和电网调度[7]。这些应用所采用的学习算法均是各自独立的, 没有考虑通过协调来优化系统的整体性能。

本文针对信息不完备的情况, 提出在各智能体独立学习的基础上建立协调策略, 以实现协作决策过程的收敛; 并以网格对策进行仿真研究, 以证明协

收稿日期: 2000-11-07; 修回日期: 2000-04-28
作者简介: 李晓萌(1975—), 男, 四川绵阳人, 博士生, 从事分布式智能控制、机器学习等研究; 许晓鸣(1957—), 男, 上海人, 副校长, 教授, 博士生导师, 从事复杂系统的智能控制等研究。

作的有效性。本文对所有智能体均采用相同的奖赏函数,即一般和对策。这样,当多智能体达到平衡点时,其联合决策也是最大折扣奖赏意义上的优化策略(当平衡点的奖赏值相同时,则为最优策略)。

2 分布式独立强化学习

在有限状态集合 $S = \{s_i\}$ 表示的问题空间中,设 $\alpha = \{1, 2, \cdots, m\}$ 为 m 个智能体的集合,所有智能体具有相同的行动集合 $A = \{a_i\}_n$。一个自主智能体在当前状态 $s \in S$ 下,依照某一行动策略 π 采取动作 $a \in A$ 得到即时奖赏 $r(s, a)$,由 $r(s, a)$ 构成集合 R。智能体根据得到的奖赏来学习定义在整个状态集合和动作集合上的数值化的评价函数 $Q: S \times A \to \mathscr{R}$ 并依据该评价函数确定修正后的行动策略。其中 $Q(s, a)$ 表达了智能体在状态 s 采取动作 a 后的期望奖赏。

强化学习的一个重要方法是 Watkins 提出的 Q-学习算法[8],这是一种动态差分算法,其一般形式为

$$Q_{t+1}(s, a) = (1 - \alpha) Q_t(s, a) + \alpha[r_t + \gamma \max_{a'} Q_t(s', a')] \quad (1)$$

其中,$Q_t(s', a')$ 为智能体在状态 s 下采取动作 a 后的下一状态 s' 的评价函数值。由智能体在所有遍历状态时的 Q 值构成智能体的 Q 值表,Q 值表的状态行动对即为下一次的行动策略。Watkins 证明了智能体无限次地遍历状态空间可以得到收敛的行动策略。本文讨论的状态变迁映射是确定性的,即智能体采取行动后到达的下一状态是确定的。该映射可表示为: $T: S \times A \to S$, $\alpha \in [0, 1]$ 为学习率, $\gamma \in [0, 1]$ 为折扣率, t 表示状态-行动对 (s, a) 的迭代次数。

对于分布式环境下的多智能体系统,智能体通常没有关于其它智能体完备的观察信息,但可假定智能体都具有最大化自己效用的理性,从而所有智能体都选择各自的最优行动策略,构成优化的联合行动。记 a^{-i} 为除智能体 i 的其它智能体的联合行动, a^u 为所有智能体的一个联合行动,则每个智能体的强化学习过程可确定为[2]

$$Q^i_{t+1}(s, a) = (1 - \alpha) Q^i_t(s, a) + \alpha[r_t(s, a) + \gamma \max_{a^{u'}} Q_t(s', a^u)] \quad (2)$$

式(2)的不足之处是在对策结构不清楚时,要求得状态 s' 下联合行动的 $a^{u'}$ 的最优值 $\max Q_t(s', a^u)$ 很困难。在信息完备的情况下,式(2)可采用求解 Nash 平衡点混合策略的方法建立迭代规则,即

$$Q^i_{t+1}(s, a) = (1 - \alpha) Q^i_t(s, a) + \alpha[r^i_t(s, a) + \gamma Q^i_t(s') \prod_{j \in \alpha} \pi^j(s')] \quad (3)$$

其中,$\pi^j(s')$ 是智能体 j 在状态 s' 的 Nash 平衡点混合策略。在一般和对策下或当智能体数目较多时,这样的混合策略不易求得。采用 Bayes 法则或假想对局的方法来估计其它智能体的策略也是常用的方法[1,3],但其缺点是不能确保协作决策过程的收敛性。

在信息不完备的情况下,智能体采取独立学习的方式。与联合学习不同,每个智能体都是一个独立的学习体,它仅利用当前状态信息、自己的行动策略和状态变迁所得到的奖赏进行学习。智能体不知道在联合行动中其它智能体的行动策略,它只维护自己的一个关于状态-行动对的 Q 值表 $Q^i(s, a)$,并且每个智能体 i 各自采取独立的迭代过程

$$Q^i_{t+1}(s, a) = \max\{(Q^i_t(s, a), r_t(s_t, a^u_t) + \gamma \max_{a' \in A} Q^i_t(s', a')\}$$
$$s_t = s, \quad a^i_t = a \quad (4)$$

由式(4)可知,迭代过程是一个单调非减的过程。假定每个智能体都认为其它智能体采取各自的优化策略,由此可得到一个优化的联合行动策略。采用该分布式学习算法的意义在于:在保证问题求解精度的前提下减少复杂问题的计算量,并由此构造可并行的计算方式。

3 智能体间的协作

每个智能体都进行各自的 Markov 决策过程(MDP),状态的变迁决定于所有智能体的联合行动,称这样的决策过程为多智能体协作决策过程(MACMDP),定义为 (S, α, A, T, R),其中各符号的意义同上。

在协作决策过程中,智能体 i 的 Q 值表中每个值都表示某个优化策略在各个状态下优化动作的奖赏值。因此,在对状态的无限次遍历后,可计算出收敛的智能体优化行动策略。首先给出智能体优化策略的更新规则

$$\pi^i_0(s) = a \in A, \quad a \text{ 任取}$$
$$\pi^i_{t+1}(s) = \begin{cases} \pi^i_t(s), & s_t \neq s \text{ or} \\ \max\limits_{a \in A} Q^i_t(s, a) = \max\limits_{a \in A} Q^i_{t+1}(s, a) \\ \arg\{\max\limits_{a \in A} Q^i_{t+1}(s, a)\}, & \text{else} \end{cases} \quad (5)$$

式(5)表示智能体仅在能改进 Q 值的情况下修

改自己的行动策略,否则维持原行动策略。设每个智能体都采用该策略,于是得到智能体协作决策的联合行动策略: $\pi^u(s) = (\pi^1(s), \cdots, \pi^m(s))$。注意, $t+1$ 步的行动策略实际上将决定 $t+2$ 步的行动。

在信息不完备的情况下,即智能体间不能交互各自的行动策略时,在某一状态下,智能体根据 Q 值表中该状态对应的状态行动对而采取动作,并根据环境反馈的奖赏来修正该状态的 Q 值和状态行动对。在智能体系统未达到 Nash 平衡点时,总有智能体能找到使自己得到更多奖赏的行动策略,直到每个智能体都不能再搜索到更优的策略。

考虑式(4)中多智能体联合行动后每个智能体所得到的奖赏都相等,并且等于环境对联合行动的奖赏,即

$$r(s,a^u) = r^i(s,a^i) = r^j(s,a^j)$$
$$\forall i,j \in \alpha,\ a^u = (a^1, \cdots, a^i, \cdots, a^j, \cdots, a^m) \quad (6)$$

式(6)表明,对智能体的联合行动采用一致的评价函数,可以确保所有智能体建立共同的目标,通过联合行动达到优化策略的平衡点。实际上,目标优化问题、多智能体网格对策的评价函数都具有上述特点。从对策论角度说,这是一种特殊的一般和对策。下面将证明基于式(6)的协作决策过程是收敛的。

引理 **1**[9]　定义联合行动的 Q 值学习过程为

$$Q_{t+1}(s,a^u) = \begin{cases} Q_t(s,a^u), & s_t \neq s \text{ or } a_t^u \neq a^u \\ r(s,a^u) + \gamma \max\limits_{a^{u'}} Q_t(s',a^{u'}) & \\ s_t = s, \quad a_t^u = a^u & \end{cases} \quad (7)$$

如果该过程无限遍历所有的状态行动对,则 Q 值将收敛到最优值 Q^*。

定理 **1**　在 MACMDP 中,如果每个智能体都采用式(4)的学习算法和式(5)的策略规则,则该 MACMDP 是一个收敛过程。

证明　定理的证明分为 3 步:

1) 分布式学习中智能体 i 的任意状态-行动对 (s,a) 的 Q 值,都是状态 s 下该智能体采取行动 a 的所有联合行动的最优 Q 值,即

$$Q_t^i(s,a) = \max_{a^i=a} Q_t(s,a^u) \quad (8)$$

由对迭代步 t 采用归纳法可以证明。

2) 对任意状态 s 和任意迭代步 $t \in \{0,1,\cdots\}$,联合行动策略都是最优的,即

$$Q_t(s,\pi_t^u(s)) = \max_{d^u \in A^u} Q_t(s,a^u) \quad (9)$$

其中, $\pi_t^u(s) = (\pi_t^1(s), \cdots, \pi_t^m(s))$ 各分量由式(5)给出。

现在分两种情况进行证明:

①　$\max\limits_{a^u \in A^u} Q_t(s,a^u) < \max\limits_{a^u \in A^u} Q_{t+1}(s,a^u)$

由式(8)可得

$$\max_{a \in A} Q_t^i(s,a) < \max_{a \in A} Q_{t+1}^i(s,a)$$

根据式(5)的策略规则,给出 $t+1$ 时刻的联合行动策略为

$$a_{t+1}^u = (\pi_{t+1}^1(s), \cdots, \pi_{t+1}^m(s))$$
$$\pi_{t+1}^i(s) = \arg(\max_{a \in A} Q_{t+1}^i(s,a)), \quad i = 1,2,\cdots,m$$

因为 Q_{t+1} 仅在 (s,a_{t+1}^u) 上进行修正,并且由各智能体的修正策略可得

$$Q_{t+1}(s,a_{t+1}^u) = \max_{d^u \in A^u} Q_{t+1}(s,a^u)$$

②　$\max\limits_{a^u \in A^u} Q_t(s,a^u) = \max\limits_{a^u \in A^u} Q_{t+1}(s,a^u)$

设 $a_t^u = (a_t^1, a_t^2, \cdots, a_t^m) = \arg\max\limits_{a^u = A^u} Q_t(s,a^u)$,由式(5)的策略规则有

$$\pi_{t+1}^i(s) = \pi_t^i(s) = a_t^i, \quad i = 1,2,\cdots,m$$

因此, $t+1$ 步的联合行动 $a_{t+1}^u = (\pi_{t+1}^1(s), \cdots, \pi_{t+1}^m(s))$,使得

$$Q_{t+1}(s,a_{t+1}^u) =$$
$$Q_{t+1}(s,(\pi_{t+1}^1, \cdots, \pi_{t+1}^m)) = \max_{a^u \in A^u} Q_{t+1}(s,a^u)$$

3) 由 2)的结论和引理 1 可以推出策略 π_t^u 将收敛到某个优化策略 π_t^{u*}。□

4　仿真研究

如图 1 所示,本文采用 5×5 的网格对策对多智能体的联合决策过程进行研究。这是一个基于一般和对策的随机对策和强化学习环境,它类似于 Sutton 和 Barto 的网格对策[10]。将智能体所处的位置称为状态,智能体的行动集合包含 4 个行动: {Right, Left, Up, Down},两个智能体分别从初始状态 (S_1, S_2) 出发,采用各自的行动策略到达目标状态 (G)。只要其中任意一个智能体到达目标即结束一轮对策,两个智能体返回各自起点重新对策。当没有智能体到达目标状态时,各获得奖赏值 -1;当至少一个智能体到达目标时,各获得奖赏值 100。智能体学习的结果是寻找从任意位置出发到达目标状态的最优策略,即最短的路径。

我们分别对单个智能体、具有完备信息的多个智能体和具有不完备信息的多个智能体进行研究。后二者的区别在于:智能体是否具有对方的对策结构和是否了解对方的学习结果(Q-值表),并根据对方的优化行动策略来改进自己的策略。其中,单个智

能体采用一般的强化学习算法, 具有完备信息的多个智能体均采用相同的最小最大强化学习算法, 具有不完备信息的智能体采用本文的强化学习算法。

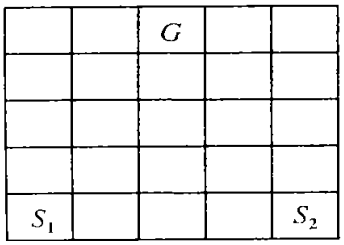

图 1 网格对策

对上述 3 种情况分别进行 10 次实验, 每次实验进行 5 000 步学习, 取 10 次中最好的一次结果。仿真结果如图 2 所示, 其中点划线、虚线和实线分别表示单个、多个完备信息智能体和多个不完备信息智能体的学习过程。以最终的收敛 Q 值表(即最优期望奖赏) 作为标准, 检查学习过程是否收敛。可以看出, 单个智能体的学习过程缓慢, 并且无法收敛到最优期望奖赏。其原因是单个智能体的行动存在盲区, 它在某一位置总是采取一种行动策略而忽视了其它可能的行动方向。多个智能体的协调恰好可以弥补这一缺陷, 并且可以得到较快的收敛结果。由于智能体之间的信息不能共享, 所以收敛速度较完备信息智能体的学习收敛速度慢, 但是已经证明, 这一学习过程同样收敛。

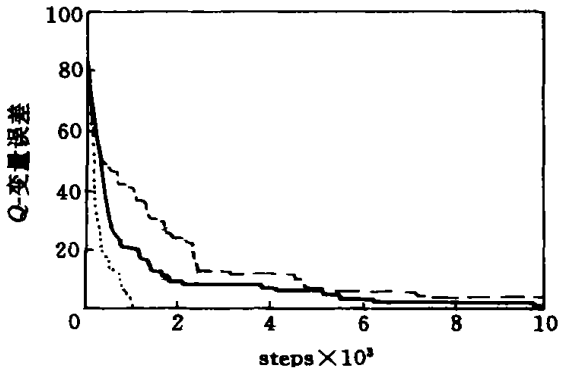

图 2 智能体强化学习

5 结 语

基于对策论和强化学习的协作决策过程, 是研究在动态环境中分布式智能控制的新方法。在智能体之间信息完备的情况下进行协作决策, 具有良好的决策性能。但对于分布式环境, 则由于时间或空间的约束, 智能体之间不能完全共享信息, 这种情况下的协作决策只能建立在智能体的独立强化学习上。理论和仿真实验都表明, 这种方法在信息有限的情况下仍可收敛到一个优化策略, 并取得比单个智能体决策更好的效果。

参考文献(References):

[1] M L Littman. Markov games as framework for multi-agent reinforcement learning[A]. Proc of the 11th Int Conf on Machine Learning[C]. San Francisco: Morgan Kaufmann, 1994. 157-163.

[2] J Hu, M P Wellman. Multiagent reinforcement learning: Theoretical framework and an algorithm[A]. Proc of the 15th Int Conf on Machine Learning[C]. Morgan Kaufmann, 1998. 242-250.

[3] C Claus, C Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems[A]. Proc of the 15th National Conf on Artificial Intelligence[C]. Cambridge MIT Press, 1997. 235-262.

[4] D H Wolpert, K Wheeler, K Tumer, et al. General principles of learning-based multi-agent systems[A]. Proc of the Third Int Conf of Autonomous Agents[C]. Seattle, 1999. 77-83.

[5] J A Boyan, M L Littman. Packet routing in dynamically changing networks: A reinforcement learning approach[J]. Adv in Neur Inform Proc Syst, 1993, 6: 671-678.

[6] R H Crites, A G Barto. Elevator group control using multiple reinforcement learning agents[J]. Machine Learning, 1998, 33: 235-262.

[7] J Schneider, W K Wong, A Moore, et al. Distributed value functions[A]. Proc of the 16th Int Conf on Machine Learning[C]. San Francisco: Morgan Kaufmann, 1999. 371-378.

[8] C Watkins. Q-learning[J]. Machine Learnning, 1992, 8: 279-292.

[9] C Watkins. Learning from delayed rewards[D]. Cambridge: Cambridge University, 1989.

[10] A G Barto, R S Sutton, C Watkins. C Learning and sequential decision making[A]. Learning and Computational Newroscience: Foundation of Addaptive Networks[C]. Cambridge MIT Press, 1990. 539-602.