基于强化学习与对策的多代理协同技术

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　本文从强化学习与Markov对策相结合方面考察了多代理协同技术的发展，系统地分析了已有的研究成果，并指出基于强化学习与对策的多代理协同技术研究中存在的问题及未来研究方向。

关键词　多代理协同，强化学习，Markov对策

Multiagent Coordination Based on Reinforcement Learning and Games

ZHANG Hua-Xiang　HUANG Shang-Teng
(Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai 200030)

Abstract　The technology of multiagent coordination based on reinforcement learning and Markov games is reviewed in this paper, and asystematic analysis of the reseach achievement in this field is introduced. Open problems of multiagent reinforcement learning and games in multiagent coordination are discribed, and future research is pointed out.

Keywords　Multiagent coordination, Reinforcement learning, Markov games

1　引言

多代理系统是分布式人工智能的主要研究领域之一，而代理协同机制是多代理系统的主要研究内容。随着代理技术的发展及应用，多代理系统中代理协同机制的研究越来越受到人们的重视。多代理系统协同机制的研究目前主要有以下几种方法：基于BDI理论模型的代理逻辑框架。将Bratman[1]的代理信念、愿望及意图理性的观点应用的MAS，提出了代理联合意图、社会规则与社会承诺、合理化行为等描述多代理行为与协同的形式化逻辑理论；基于协商的多代理协同机制[2]。代理通过协商来消解冲突、分配资源及联合求解任务。合同网(Contract Net)[6,7]是基于代理协商的；Gensereth[4]与Gmytrasiewicz[8]将对策理论应用到多代理合作，讨论了无通信情况下代理间的协作；基于联合规划的代理协同。如全局规划及部分全局规划[9]，通过协调代理的目标实现代理协同；基于组织与社会规则的代理协同。如基于联盟的代理组织[10,11]。以上代理协同机制都是在系统设计时事先定义好。对于确定的系统模型，离浅定义代理的协同机制可以提高协同效率，但缺乏灵活性及可扩展性，尤其不适用于动态系统。当系统的模型未知时，不能准确地离线定义代理的协同机制。为此，提出了适用动态、未知模型系统的自协调模型。多代理动态系统中，代理通过与环境不断交互，调整自身行为，来实现协同。多代理的自协调模型主要通过动态环境下代理的分布式学习实现。系统通过效用函数对代理的行为进行评价。通过学习，代理可以实现自身行为策略的优化，从而实现效用的最大化。由于基于学习的代理协同机制具有很好的可扩展性、鲁棒性，已广泛地应用于机器人足球、交通控制、电子商务、网络管理、信息检索、计算机辅助教学与医疗、分布式计算、多处理器协同设计等领域。

2　基于学习与对策的多代理协同研究方法

Markov对策(MG)[12]是多代理协同技术研究的基本模型之一。MG将Markov决策理论应用到对策论，将单阶段对策扩展到多阶段，且各阶段间具有Markov属性。MG中每个代理的行为是不同的Markov决策过程。代理通过行动的收益来评价行动的好坏。MG以Nash均衡点为代理协作的目标；基于学习的代理协作。代理建立其它代理的模型并通过学习修正该模型实现自身利益最大化。学习方法主要有虚拟回合、贝叶斯信念学习及强化学习。虚拟回合中[13]代理根据其它代理执行不同联合行动的概率来选择最优的行动策略。贝叶斯信念学习指代理根据贝叶斯法则不断地修改初始建立的对手代理行动信念。强化学习是指代理 通过观察不同行动从环境中得到的强化信号对行动进行评估。通过不断的试错，最终得到最优策略；基于决策与对策树的多代理协同。对策论没有给出代理如何获得最优策略。Cristina[14]将对策论和对策过程形式化，通过描述决策树上代理的信息集合与行为函数，推理行动策略。对策全过程用对策树描述，树中分支表示不同的选择，节点表示知识及状态，叶节点表示收益。该方法适合于有限状态与有限行动集的对策问题求解；基于决策网络的代理协同。Gmytrasiewicz[15]以决策网络形式描述动态MAS，主要思想是用图的形式描述决策问题。图中有三种节点分别为自然节点、决策节点及收益节点，节点间的连接表示节点间存在依赖关系。代理通过与环境交互，不断修改初始信念、偏好及能力的描述，直至稳定。

3　强化学习

早期将学习应用于多代理协同基于先验知识[6]和代理共享合作域知识[17]，如代理通讯合作共享知识的研究[18~20]。

强化学习[21,22]近年来被越来越多地应用到多代理领域。最早见于Tan、Sen、sandholm、weiss[18,23~25]等人的工作，他们将强化学习的方法直接应用到MAS。随后基于Q学习[26,27]提出的各种多代理强化学习算法主要有Littman的minimaxQ学习[28]、Hu和Wellman的NashQ学习[29,30,34]、Claus和Boutilier[31,32]的全合作多代理Q学习、Bowling和Veloso

的可变学习率Q学习[33,35,36]、Littman的FFQ学习[37]及Greenwald的关联均衡Q学习[38]。

3.1 Markov决策(MDPs)

代理某一时间选择行动,从环境中得到回报并以一定的概率转移到新状态。该过程称为一个决策过程,若代理的状态转移具有Markov性质,则称该决策过程为MDPs。

动态规划[39]是求解最优化问题的技术之一。给定MDps的状态转移概率和收益函数,代理最优策略可通过求解Bellman最优化方程得到。但系统模型未知时,动态规划不再适用。

3.2 单代理强化学习

强化学习是一种无模型的学习。通过对状态行动的随机搜索,最终实现函数逼近。与动态规划有所不同,TD学习根据一步或多步行动后从环境中得到的回报对当前状态值函数评估。如果代理根据一步行动回报评估状态值,称为TD(0);否则称为TD(λ)。如果代理直至行动结束才对以前各步行动进行评估,则称为蒙特卡洛(monte carlo)方法。

Dayan[40]证明当学习率以一定方式衰减,且所有状态被无限次地遍历,TD(0)学习以概率1收敛。TD(0)与策略有关。Q学习[26]是一种与策略无关的TD学习,当状态行动对被无限次地遍历后,Q值以概率1收敛到最优,且与更新策略无关。

4 Markov对策与多代理强化学习

对策论[42]是研究自利代理间相互交互的形式化理论。主要研究代理在交互过程中如何通过选择策略来最大化自身利益。对策论应用于MAS的研究主要是通过Rosenschein等人的工作[43~46]。通过引入对策,在一定的理性假设前提下,代理可以在无通讯情况下实现协同。

将状态的概念引入到对策中,若状态转移具有Markov性质,则称为MG。MG是一个五元组$(I,S,\{A_i\}_{i\in I},P,\{R_i\}_{i\in I})$。分别表示代理集合、状态集合、代理可选行动集合、状态转移概率函数及代理的立即收益。MG的混合策略表示代理在其行动集合上的一种概率分布。对策中所有代理的混合策略组合构成对策的混合策略。带折扣MG的目标是最大化代理的累计折扣收益。针对MG有结论[12]:具有稳定策略的任一带折扣的随机对策至少存在一个Nash均衡点。

4.1 对策中的学习

对策论很多工作集中在如何获取对策。如虚拟回合[13]、对手建模[49,31]、学习自动机[51]、贝叶斯信念学习等。虚拟回合中,代理计算自身行动的平均值。并选择使平均值最大的行动。虚拟回合在全合作对策中可以收敛到Nash均衡点[31],同时对于零和对策可以得到基于经验分布的经验策略。基于对手建模[31,49],代理推测其它代理采取的联合行动,并据此概率选择最优策略。Claus和Boutiler[31]研究了全合作对策中的对手建模,Uther和Veloso[49]研究了全竞争对策中的对手建模,Hu和Wellman[50]提出了代理递归建模的方法。学习自动机基于代理独立学习并不断更新自身行动的概率分布。文[51]给出结论:对于重复对策,代理以概率1收敛到一个均衡点。Verbeeck[52]利用学习自动机思想提出代理通过学习达到最优均衡点的算法。

4.2 基于强化学习的对策选择

代理通过学习Q值函数进行决策,其Q学习规则为

$Q_i(s,a_i)=(1-\alpha)Q_i(s,a_i)+\alpha(R_i(s,\vec{a})+\gamma V_i(s'))$,$V_i(s)=\max_{a_i\in A_i}Q_i(s,a_i)$

MAS中代理收益受到代理联合行动的影响,于是有

$Q_i(s,\vec{a})=(1-\alpha)Q_i(s,\vec{a})+\alpha(R_i(s,\vec{a})+\gamma V_i(s'))$,其中$\vec{a}$为代理的联合行动

上式有一个基本的假定:系统中代理所能采取的行动是已知的,即系统中所有代理的行动可观察。针对不同的情况,$V_i(s)$各不相同。

(1)零和MG

两代理零和对策中代理收益之和为零。若代理1试图最大化收益,则代理2试图最小化该收益。Q值的更新规则为

$$Q_1(s,a_1,a_2)=(1-\alpha)Q_1(s,a_1,a_2)+\alpha(r_1+\gamma \max_{\pi_1(s,a_1)\in\Pi(a_1)} \min_{a_2'\in A_2}\sum_{a_1'\in A_1}\pi_1(s,a_1)Q_1(s',a_1',a_2'))$$

Littman等[48]证明在零和两代理环境中,按照minimaxQ的更新规则学习的Q值函数收敛到最优w.p.1。如果代理执行GLIE[48]的搜索策略,且均衡唯一,则代理行动收敛w.p.1。

(2)一般和MG

Hu和Wellman提出了针对一般和的NashQ学习算法。更新规则为

$$Q_i(s,\vec{a})=(1-\alpha)Q_i(s,\vec{a})+\alpha(R_i(s,\vec{a})+\gamma V_i(s')),$$

$$V_i(s)=NashQ_i(s)=\sum_{\vec{a}}\pi_1^*(s,\vec{a})\pi_2^*(s,\vec{a})\cdots\pi_i^*(s,\vec{a})\cdots\pi_n^*(s,\vec{a})Q_i(s,\vec{a})$$

$V_i(s)$为单阶段以$\{Q_i(s),i=1,2,\cdots,n\}$为对策的Nash均衡点处第$i$个代理的立即收益。

可以看出minimaxQ学习是NashQ学习的一种特殊情况。Hu等[29,30,34]证明多代理环境中代理按照NashQ学习更新Q值,若Q值函数满足一定的条件[30],则Q值函数收敛到最优w.p.1。如果代理执行GLIE的搜索策略,且均衡点唯一,则代理行动收敛w.p.1。

一般情况下,一般和对策中有多于一个的Nash均衡点。因此NashQ学习不适合更一般的情况。Littman将NashQ学习解释为FF-Q算法[37],并证明了Q值函数的收敛性。

实际应用中NashQ学习有很多受限的地方。一方面不能保证代理选择相同的均衡点;另一方面求解均衡点需要用到二次规划,计算复杂。

Greenwald基于对策论中关联均衡的思想,提出了关联Q学习[38]。此时

$$V_i(s)=CE_i(Q_1(s,\vec{a}),Q_2(s,\vec{a}),\cdots,Q_n(s,\vec{a}))=\sum_{\sigma\in A}\sigma^*(\vec{a})Q_i(s,\vec{a})$$

求解关联Q学习均衡点需要线性规划的方法。Greenwald证明了算法的收敛性,并说明关联Q学习算法符合Bowling[33]提出的多代理学习收敛性与理性两条件。

(3)全合作MG

全合作对策中,代理具有相同的收益函数。此时Q值更新只需要对一个代理进行即可。Szepesvari和Littman[48]证明具有相同收益函数的多代理合作Q学习的收敛性。针对重复全合作对策,Claus和Boutiler[31]给出基于联合行动学习(JAL)(Joint Action Learning)和代理独立学习IL(Independent Learning)的Q值更新规划。

5 最新研究及存在的问题

5.1 全合作对策多代理学习

多代理全合作对策研究的主要问题是代理如何通过学习

协同行动，使代理的收益达到最优。

代理全合作对策[31]中，系统模型未知时，代理需要通过学习实现合作。Claus 和 Boutilier[31,32]的 IL 学习基于对手行动不可知假设，代理学习自身行动 Q 值。JAL 基于对手行动可观察假设。代理学习联合行动 Q 值决定行动策略。实验表明，两种代理学习都行动收敛，但不能保证收敛到最佳行动。

基于乐观假设，Lauer 和 Riedmiller[58]提出代理 IL 学习搜索最优行动的算法。乐观假设认为代理都采用各自的最优策略，从而构成代理的最优联合策略。基于乐观假设的代理学习算法 Q 值收敛，且行动收敛到最优行动。

Wang 和 Sandholm[55]提出了最优适应学习算法。该方法基于 Young[3]提出的适应游戏(adaptive play)思想。算法能保证代理收敛到最优联合行动。

Guestrin 等[59]提出部分可视信息下，代理通过函数估计协调行动。作者应用最小平方策略迭代[60]，对 Q 函数使用线性函数估计。通过与环境交互得到训练数据，再使用训练数据估计参数。该方法能保证参数的收敛，并保证 Q 值收敛。

Mukherjee 和 Sen[53]提出代理通过学习实现 pareto 最优的方法也适用于全合作对策。基于少量通讯和 IL，代理交替或同时揭示行动，并在对方行动已知时选择最优行动。实验表明多 Nash 均衡点存在时，针对一定结构的一般和对策，代理揭示行动增加收敛到最优 Nash 均衡的机会。

针对全合作对策中寻找 pareto 最优 Nash 均衡点，Verbeeck 等[52]提出不断寻找对策中的 Nash 均衡点，并保留最大的。一定数量回合的学习后，代理一定能够学习到 pareto Nash 均衡。

Chalkiadakis 和 Boutilier[56]提出了全合作多代理强化学习 Bayesian 方法。代理利用 Bayesian 规则对系统模型信念及对手策略信念进行推理。针对合作对策，实验结果表明该方法可以提高在线学习效率，且总能收敛到最优策略。其收敛速度快于 Bowling[36]的 WoLF 学习。

随机环境和半随机环境下代理全合作对策由 Kapetanakis 和 Kudenko[57]提出。有微小干扰的情况下 Wang 和 Sandholm[55]的最优适应学习针对惩罚对策也能收敛到最优行动。Lauer 和 Riedmiller[58]的学习策略可能导致错误的行动收敛。其它针对确定环境的方法皆不适用。Kapetanakis 等[57]提出一种启发式搜索策略。实验结果表明在半随机环境下代理使用该策略可收敛到最优行动，但不适合随机环境。上述针对全合作对策代理学习最优行动的算法都是针对于确定性对策模型，不适合随机及半随机域中的代理协同[57]。

5.2 一般和对策多代理 Q 学习

零和随机对策及全合作随机对策都可以看成一般和随机对策的特殊情况。因此若能从理论上很好地解决一般和随机对策的多代理学习，其它情况可作为特例处理。Shoham 和 Powers[54]讨论了多代理强化学习，指出未来需要解决的几个方面的问题。

多代理环境中代理行动收益为代理联合行动的函数，因此一般和随机对策中代理强化学习都是针对联合行动的 Q 值学习。通过与其它代理的交互，代理学习最佳的行动策略。两代理零和对策下 Littman[28]的 minimax-Q 学习基于悲观假设(假定对手的行动最小化代理的收益)，代理学习最差情况下的最优策略。Hu 和 Wellman 的 Nash-Q 学习策略基于对手理性行动假设，代理的行动策略为选择每一个阶段对策的 Nash 均衡点，并假设对手亦选择 Nash 均衡点策略。当有多个 Nash 均衡点存在时，一方面需要确定代理如何选择最优的均衡点，同时要确定如何保证所有代理选择相同的 Nash 均衡点。

代理学习收敛要求每个阶段对策都存在严格的最优点或鞍点作为 Nash 均衡点，而实际问题很难满足上述条件。两代理网格对策的实验结果[30,47]表明，阶段对策只有一个 Nash 均衡点存在时，Q 值函数收敛。阶段对策有多个 Nash 均衡点存在时，Q 值函数不收敛。基于对手选择 Nash 均衡点的假设，当有一个均衡点存在时，代理的 Q 值收敛。

Littman[37]的 FF-Q 学习基于乐观和悲观两种极端假设，该算法能保证收敛。

Greenwald[38]提出的基于关联 Nash 均衡点的 Q 学习(CE-Q)，其基本思想类同于 Nash-Q 学习，只是代理在 Q 学习更新规则中选择关联 Nash 均衡点处代理的单阶段对策期望收益。

CE-Q 学习满足 Bowling[33]提出的理性和收敛性两条件，当均衡点多于一个时依然存在选择均衡点的问题。存在单均衡点时基于对手同样选择均衡点策略假定下，CE-Q 学习收敛。若对手采取不同的学习策略，比如 Nash-Q 学习、Q 学习及 FF-Q 学习时，情况如何？

Suematsu 和 Hayashi[5]针对一般和对策问题提出了一种扩展式最优响应学习。代理记录对手行动的策略、代理行动，并观察新的状态及对手的行动，更新对手行动策略信念，同时按照一定的规则更新自身原来状态的行动策略。

以上基于假设的多代理学习，代理需要联合行动信息，同时需要所有代理的每个对策阶段的 Q 值，以便计算均衡点。这样做有两方面的缺点：需要大量的存储空间；增加代理间的通讯。为此需要有新的基于对手行动策略假设的代理学习方法，减少代理对 Q 值的存储(即代理只要存储自身的 Q 值)，同时基于对手行动策略的假设，代理又有最优的学习策略。

结束语 协同技术作为多代理系统的重要研究内容之一，受到了越来越多的重视。随着代理经济的出现与发展，具有可扩展性、鲁棒性的适用于动态多代理系统的协同技术必将成为跨学科的新的研究领域。目前基于学习与对策的多代理协同的研究刚刚起步，各种学习规则都是基于代理理性假设，多代理学习还没有统一的理论框架。

参 考 文 献

1 Bratman M E. Intention, Plans and Pratical Reason. Harvard University Press, Cambridge, MA, 1987

2 Conry S E, Meyer R A, Lesser V R. Multistage negotiation in distributed planning. Readings in distributed artificial intelligence, morgan kaufman, 1998. 367～384

3 Young H. The evolution of conventions. Econometrica, 1993, 61(1): 57～84

4 Gensereth M, Ginsberg M, Rosenschein J. Cooperation without communications. In: Proc. of the national conf. on artificial intelligence. Philadelphia, Pennsylvania. 1986. 51～57

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

6 Smith R G. The Contract-Net Protocol: High Level Communication and Control in a Distributed Problem Solver. IEEE Trans. on Comp, 1980, C29(12): 1104～1113

7 Sandholm T W, Lesser V R. Coalition among computationally bounded agents. Artificial Intelligence, 1997. 99～137

8 Gmytrasiewicz P J, Durfee E H. Rational Coordination in Multi-

aggent Environments. Autonomous Agents and Multi-agent Systems,2000,3:319～350
9 Durfee E H,Lesser V. Negotiating Task Decomposition and Aplication Using partial Global Planning. Distributed Artificial Intelligence,Pitman Publishing:London and Morgan Kaufman:San Mateo,CA,1989,2:229～244
10 Shehory O,Kraus S. Formation of overlapping coalitions for precedecce-ordered task execution among autonomous agents in :ICMAS-96,330～337
11 Shehory O,Krars S. Methods for task allocation via agent coalition formation. Artificial Intelligence,1998,101:165～200
12 Filar D,Vrieze K. Competitive Markov Decision Process. Springer-Verlage,1997
13 Vrieze O J. Stochastic games with finite state and action spaces [M]. CWI Tracts,1987(33)
14 Cristina B,Eithan E,et al. Games Servers play:A procedual Approach
15 Dicky S,Piotr J G. Learning Models of Other Agents Using Influece Diagrams
16 Brazdil P,et al. Learning in distributed systems and multi-agent environments. In European working session on learning ,lecture notes in AI,482,Berlin. Springer Verlag,1991
17 Sian S. Adaptation based on cooperative learning in multi-agent systems. Decentralized AI,Elsvier science publications,1991,2:257～272
18 Tan M. Multi-agent reinforcement learning:Independent vs. Cooperative agents. In:proc. of the tenth intl. conf. on machine learning. 1993. 330～337
19 Prasad M V N,Lander S E,Lesser V R. Cooperative learning over composite search spaces:experiences with a multi-agent design systm. In:Proc. of thirteeth national congerence on atficial intelligence. 1996. 68～73
20 Provost F J,Hennessy D N. scaling up:distributed machine learning with cooperation. In:Proc. of the thirteeth national conf. on artificial intelligence. CA. AAAI press,1996. 74～79
21 Sutton R S,Barto A. Reinforcement Learning:An Introduction [M],MIT press,Cambridge,MA,1998
22 Kaelbling L,Littman M L,Moore A W. Reinforcement leaning:A surey. Journal of Artificial Intelligence Research,1996,4:237～285
23 Weiss G. Learning to coordinate actions in multi-agent systems. In IJCAL,1993
24 Sen S,Sekaran M,Hale J. Learning to coordinate without sharing information. In AAAL,1994
25 Sandholm T,Crites R. Learning in the iterated prisoner's dilemma. Biosystems,1995,37:147～166
26 Watkins C J C H. Learning from Delayed Rewards:[Ph. D. thesis]. Cambridge,UK:Cambridge University
27 Watkins C J C H, Dayan P. Q-learning. Machine Learning 1992,8:272～292
28 Littman M L. Markov games as a framework for multi-agent reinforcement learning. In:11th ICML,New Brunswick,1994. 157～163
29 Hu junling,Wellman M P. Multiagent reinforcement leaning:Theoretical framework and an algorithm. In:15th ICML,p242～250
30 Hu Junling, Wellman M P. Nash Q-Learning for General-Sum Stochastic Games. Journal of Machine learning research,2003(1):1～30
31 Claus C,Boutilier C. The dynamics of reinforcement learning in cooperative multiagent systems. In:proc. of the Fifteenth National Conf. on Artificial Intelligence,1998
32 Craig Boutilier Sequential optimality and coordination in multiagent systems. In:16th Intl. Joint Conf. on Artificial Intelligence,Stockholm,1999. 478～485
33 Bowling M, Veloso M. Rational and converagent learning in stochastic games. In:17th Intl. Joint Conf. on Artificial Intelligence,2001. 1021～1026
34 Bowling M. Convergence problems of general-sum multiagent reinforcement learning. In:Proc. 17th ICML,Stanford,CA,Morgan Kaufmann,San Francisco,CA,2000. 89～94
35 Bowling M,Veloso M. Variable learning rate and the convergence of gradient dynamics. In:Proc. 18th ICML,Williamstown,MA,2001. 27～34
36 Bowling M,Veloso M. Multiagent learning using a variable learning rate. Artificial Intelligence,2002,136:215～250
37 Littman M L. Friend-or-foe Q-learning in general-sum games. In:18th ICML,Willianis college,MA,2001. 322～328
38 Greenwald A,Hall K,Serrano R. Correlated-Q learning. In:NIPS Workshop on Multiagent Learning,2002
39 Bellman R E. Dynamic Programming. Princeton,NJ:Princeton University Press,1957
40 Dayan P. The convergence of TD(λ) for general λ. Machine Learning,1992,8:341～362
41 Rummery G A,Niranjan M. On-line Q-learning using connectionist systems:[Technical Report CUED/F-INFENG/TR 166]. Engineering Department,Cambridge University
42 Osborhe M J,Rubinstein A. A course in game theory. The MIT Press Cambridge,Massachusetts London,Englind,1994
43 Rosenschein J S. Rational interaction:Cooperation among intelligent agents:[Ph. D. thesis]. Computer Science Department,Stanford University:Stanford,CA,1985
44 Rosenschein J S,Genesereth M R. Deals among rational agents. In:Proc. Ninth int. Joint Conf. Artificial Intelligence(IJCA-85),Los Angeles,CA,1985. 91～99
45 Rosenschein J S,Ginsberg M,Genesereth M R. Cooperation without communication. In:Proc. Fifth National Conf. Artificial Intelligence(AAAI-86),Philadelphia,PA,1986
46 Rosenschein J S,Zlotkin G. Rules of Encounter:Designing Conventions for Automated Negotiation among Computers. The MIT press:Cambridge,MA,1994
47 Hu Junling,Wellman M P. Experimental Results on Q-Learning for General-Sum Stochastic Games. ICML2000. 407～414
48 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
49 Uther W,Veloso M. Adversarial reinforcement learning:[Technical Report]. Cornegie Mellon University,1997
50 Hu Junling ,Weliman M P. Learning about other agents in a dynamic multiagent system. Journal of Cognitive Systems Research,2001,2:67～79
51 Narendra K,Thathachar M. Learning Automato:An Introduction. Prentice-Hall,1989
52 Verbeeck K,Nowe A,Lenaerts T,Parent J. Learning to reach the Pareto Optimal Nash Equilibrium as a Team. Lecture Notes in Artificial Intelligence 2557. In:Proc. of AI02. Canberra,Australia,2002
53 Mukherjee R,Sen S. Towards a Pareto-optimal solution in general-sum games. In:Learning Agents Workshop Notes,5th Conf. Autonomous Agents,2001
54 Shohm Y, Powers R, Grenager T. Multi-Agent Reinforcement learning:a Manifesto. AAAI2003
55 Wang XiaoFeng,Sandholm T. Reinforcement Learning to Play an Optimal Nash Equilibrium in Team Markov Games. NIPS2002,Vancouver,Canada
56 Chalkiadakis G,Boutilier C. Coordination in Multiagent Reinforcement Learning:A Bayesian Approach. In:2nd Intl. Conf. on Autonomous Agents and Multiagent systems(AAMAS-03)2003,to appear
57 Kapetanakis S,Kudenko D. Reinforcement learning of coordination in cooperative multi-agent systems. In:Proc. of the Nineteenth National Conf. on Artificial Intelligence(AAAI). 2002
58 Lauer M,Riedmiller M. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In:Proc. of the Severenth Intl. Conf. in Machine Learning,2000
59 Guestrin C, Lagoudakis M, Parr R. Coordinated Reinforcement Learning. AAAI Spring Symposium,Stanford,California,March 2002
60 Lagoudakis M,Parr R. Model free least square's policy. iteration. In NIPS-14,2001