文章编号：1000－8152(2003)03－0317－06

多智能体系统中的分布式强化学习研究现状

仲 宇[1]，顾国昌[1]，张汝波[1,2]
（1.哈尔滨工程大学 计算机科学与技术学院，黑龙江 哈尔滨 150001；
2. 中国科学院 沈阳自动化研究所 机器人学重点实验室，辽宁 沈阳 110015）

摘要：对目前世界上分布式强化学习方法的研究成果加以总结，分析比较了独立强化学习、社会强化学习和群体强化学习三类分布式强化学习方法的特点、差别和适用范围，并对分布式强化学习仍需解决的问题和未来的发展方向进行了探讨.

关键词：分布式强化学习；多智能体系统；机器学习
中图分类号：TP18 **文献标识码**：A

Survey of distributed reinforcement learning algorithms in multi-agent systems

ZHONG Yu[1], GU Guo-chang[1], ZHANG Ru-bo[1,2]
(1. Computer Science and Technology College, Harbin Engineering University, Heilongjiang Harbin 150001, China;
2. Robotics Laboratory, Shenyang Institute of Automation, Chinese Academy of Sciences, Liaoning Shenyang 110015, China)

Abstract: The achievements in distributed reinforcement learning algorithms are introduced. The present distributed reinforcement learning algorithms are classified into three categories, and the characteristics and application ranges of three classes of distributed reinforcement learning algorithms are compared. At the end of the article, the problems to be solved and the futher works are discussed.

Key words: distributed reinforcement learning; multi-agent system; machine learning

1 引言(Introduction)

强化学习(reinforcement learning, RL)的思想来自于对动物学习过程的长期观察和巴甫洛夫条件反射学说，其概念最早于1954年由Minsky提出，20世纪80年代成为人工智能研究中的活跃领域. 瞬时差分算法(temporal differences, TD)[1]和Q学习算法[2]的提出，是强化学习理论和应用研究的一个里程碑，它们使得强化学习理论真正成熟并得到广泛应用. 强化学习的主要算法有TD算法、Q学习算法、自适应启发评价算法(adaptive heuristic critic, AHC)等，文献[3]论述了对这些算法的形式化分析及应用领域和成果.

标准的强化学习是以一个单独的学习系统为载体进行的，从智能体理论的角度来看，这个学习系统可以看作一个智能体(agent)，这个agent在感知到当前环境状态的条件下通过对环境采取试探行为，获得环境反馈的此动作对此状态适应程度的评价和新的环境状态，从而不断修改从状态到动作的映射策略. 强化学习是一个人工智能范畴中的概念，如果引申到分布式人工智能范畴，相应地就得到分布式强化学习的概念，如果把agent看作强化学习的主体，那么把多agent系统看作分布式强化学习的主体就是很自然的了. 分布式强化学习并没有一个公认的定义，笔者认为，如果强化学习系统由多个学习单元组成，每个单元独立地执行部分或全部的强化学习任务，最后达到整个系统意义上的学习目标，这个学习系统就可以称作分布式强化学习系统，这个学习过程就可以称作分布式强化学习过程.

本文全面介绍了多agent系统中分布式强化学习的主要方法，以及这些方法各自的特点和应用范围，最后讨论了分布式强化学习需要进一步研究的问题.

2 分布式强化学习研究现状及主要方法(Research status and main algorithms of distributed reinforcement learning)

多agent系统中的强化学习过程可以是集中式的，也可以是分布式的. 集中式强化学习通常把整个多agent系统的协作机制看成学习的目标，承担学习任务的是一个全局性的中央学习单元，这个学习单元以整个多agent系统的整体状态为输入，以对各个agent的动作指派为输出，采用标准的强化学习方法进行学习，逐渐形成一个最优的协作机制. 集中式强化学习系统中的各个agent都是“傻”agent，它们不能执行学习任务，而只能被动地执行学习结果. 集中式强化学习

收稿日期：2001－11－13；收修改稿日期：2002－11－20.
基金项目：中国科学院机器人学开放研究实验室基金(RL2000106)；某国防基础研究基金资助项目.

通常用于调度问题,例如电梯组调度问题[4]和异构机器调度问题[5]等.

与集中式强化学习相对应,分布式强化学习系统中的各个agent都是学习的主体,它们分别学习对环境的响应策略和相互之间的协作策略.现有的分布式强化学习算法根据系统中的agent进行决策时是否考虑其他agent的动作可以分为两大类,分别命名为独立强化学习(reinforcement learning individually,RLI)和群体强化学习(reinforcement learning in groups,RLG).RLI方法不采用组合动作(joint-action),其agent仅凭自己的状态输入来决定采取的动作,可以看作是一种松散耦合的分布式强化学习系统;而RLG方法采用组合动作,每个agent都必须考虑到其他所有agent将要采取的动作才能决定自己采取的动作,可以看作是一种紧密耦合的分布式强化学习系统.另外,目前将RLI与社会模型或经济模型的结合以获得更高智能水平的多agent系统的研究十分活跃,已经形成了一个单独的研究方向,所以将这一类分布式强化学习方法单独分离出来称为社会强化学习(reinforcement learning socially,RLS).以下将分别对这些方法进行讨论.

2.1 独立强化学习(Reinforcement learning individually,RLI)

RLI中每个agent的学习过程都不依赖于其他agent,RLI系统的agent可能只考虑自己的状态和动作而不关心其他agent的状态和动作,各个agent从信度分配模块获得的强化信号也只与自己的状态和动作相联系,所以进行学习时甚至可能忽视其他agent的存在,认为只有自己在学习.由于RLI系统中的agent都是以自我为中心的,所以不易达到全局意义上的最优目标;但是RLI系统中agent的独立性较强,容易动态增减agent的个数,而且agent个数的变化对学习收敛性的影响较小,所以RLI方法较适用于agent个数较多的复杂系统.

由于RLI系统的agent不考虑其他agent的状态和动作,所以通信是实现agent之间协作机制的主要方法,agent可以通过通信来共享传感器信息和传递强化信号等,从而实现多个agent的相互协作,实现整体学习目标[6~9].在某些任务背景下,有可能无需通信就可以实现多agent的相互协作,但在这种情况下,agent认为只有自己在与环境打交道,对同伴的存在一无所知,这样此agent就处于一种动态环境之下,同伴的行为将使得环境的变迁产生不确定性,这样就需要花费更多的学习时间,而且学习结果只适用于学习时所处的场景,所以这种无通信的盲目方法只能用于一些比较简单的问题[10,11].

一般情况下,多agent系统中交换的信息主要是传感器信息、经验序列、策略三种.共享传感器信息、共享经验序列、共享策略三种方案的通信量依次递减,收敛速度却依次递增[12].但是共享策略的缺点在于不利于保持学习的多样性,可能由于学习过程的搜索范围狭窄而使系统陷于局部极值,而共享传感器信息方案由于各个agent的独立性最强,所以学习多样性最强,能够更大范围地探索学习空间.因此,分布式强化学习系统中各agent之间应该共享哪些信息需要根据实际任务的特点仔细权衡.

还有一些RLI系统的学习过程是通过多agent之间的竞争实现的[13~17].这些竞争型的RLI系统的共同思想是每个时间步仅选择一个agent执行其动作,相应地,在此时间步内获得的强化信号归被选中的agent所有.竞争型RLI方法的着眼点和优点在于产生学习结果的多样性,使用多个agent同时学习,选择学习结果最好的agent执行其动作,从而扩大了学习过程的搜索范围,避免系统陷于局部极值的情况.竞争型RLI方法的另一个优点是结构信度分配比较容易.

RLI方法由于思想较简单,所以成为最常用的分布式强化学习方法,我国也进行了这方面的研究.中国科学技术大学的蔡庆生等人在机器人足球的任务背景下探讨了agent团队的强化学习,分析了足球任务中状态空间和动作空间的表示方法,控球队员作为主导agent进行学习,并通过主导agent的角色变换实现整个团队的学习[18].

2.2 社会强化学习(Reinforcement learning socially,RLS)

RLS可以看作是RLI的推广,是RLI与社会模型或经济模型的结合.RLS模拟人类社会中人类个体之间的交互过程,建立社会模型或经济模型,用社会学和管理学的办法来调节强化学习agent之间的关系,形成通信、协作、竞争机制,从而获得比RLI系统更大的灵活性,可以建立比RLI系统更复杂的结构,同时可以更有效地克服RLI方法自私的缺陷以及更有效地解决结构信度分配问题.

为了将较好的策略在多agent社会中传播,可以采用前面提到的共享经验序列或策略的主动传播方式[19],也可以采用经验不丰富的agent模仿经验丰富的agent的被动传播方式.Yamaguchi的方案中,agent的性能定义为最近100步获得强化值的总和,性能较低的agent模仿性能较好的agent的行为,模仿的概率是二者性能差的随机变量,agent根据与其他agent的性能差自动在模仿和独立学习之间切换[20].Price等人也提出了一种通过模型抽取使得agent模仿其他性能较好的agent的机制[21].但是这些方法只是部分地解决了向谁学习和学习什么的问题,因为经验丰富或者获得较多强化值的agent未必具有较优的策略,也可能性能较低的agent的策略的某一部分比性能较高的agent的策略更优,所以多agent社会中的策略传播还需要进一步的研究工作.

多个agent在执行任务的过程中,由于各自目标不同,资源有限,以及相互之间的其他影响,不可避免地会产生某些冲突,这就需要冲突消解机制来协调各个agent的行为.谈判[22~24]、法律[25]、挫折[26,27]等概念已用于一些冲突消解算法,另外联盟、表决、帮助、责备、夸奖等概念如何用于冲突消解还有待于进一步的研究.

针对竞争型学习系统的结构信度分配问题,RLS引入了"拍卖"或者"交易"的经济模式."拍卖"方法的一般过程是:agent通过拍卖的办法交换执行权,报价最高的agent获得本轮的执行权并向交出执行权的agent支付所报的费用,同时可以获得本轮的强化信号和下一轮的拍卖所得作为收入,破产的agent将被淘汰[28,29]."交易"方法的一般过程是:agent

从其他 agent 得到资源或服务,必须向提供资源或服务的 agent支付费用,提供后勤支持的 agent 通过这种方式就可以间接地从直接获得强化信号的 agent 分得自己应得的那部分强化信号[30]."拍卖"或者"交易"的结构信度分配方法体现了"不付出就没有回报"和"优胜劣汰"的思想,避免了不劳而获的现象,分配结果比较公平,是两种很有发展前途的结构信度分配方法.

2.3 群体强化学习(Reinforcement learning in groups, RLG)

RLG 将所有 agent 的状态或动作看作组合状态或组合动作,每个 agent 的 Q 函数表都是组合状态和组合动作到 Q 值的映射.RLG 系统中的每个 agent 都必须考虑其他 agent 的状态,选择动作时也必须考虑其他 agent 将要执行的动作,所以具有状态空间和动作空间庞大的特点,学习速度很慢,这种"紧密耦合"的方法一般只适用于 agent 很少的情况下,而且需要加速算法的支持.

Junling Hu 等人于 1998 年提出的 RLG 算法可以说是最经典的 RLG 算法,此算法用 Markov 竞赛理论作为分布式强化学习的理论框架,并证明此算法能够收敛到 Nash 均衡策略[31].文中以两个 agent 为例的 RLG 算法的 Q 值更新规则如下:

$$Q^1_{t+1}(s,a^1,a^2) =$$

$$(1-\alpha_t)Q^1_t(s,a^1,a^2)+\alpha_t[r^1_t+\beta\pi^1(s')Q^1_t(s')\pi^2(s')], \quad (1)$$

$$Q^2_{t+1}(s,a^1,a^2) =$$

$$(1-\alpha_t)Q^2_t(s,a^1,a^2)+\alpha_t[r^2_t+\beta\pi^1(s')Q^2_t(s')\pi^2(s')]. \quad (2)$$

其中 $Q^k_t(s,a^1,a^2)$ 代表第 t 个时间步,环境状态为 s 的情况下,两个 agent 分别选择动作 a^1 和 a^2 时第 k 个 agent 的 Q 值 ($k=1,2$);策略 $\pi^k(s')$ 是在状态 s' 下第 k 个 agent 的所有动作被选中的概率分布向量 ($k=1,2$);$Q^k_t(s')$ 是以 a^1 为行下标,a^2 为列下标的矩阵,其第 a^i_t 行 a^2_j 列的元素为 $Q^k_t(s',a^1_t,a^2_j)$.

本算法中每个 Q 表有 $|S|^n\times|A|^n$ 个(s,a^1,a^2) 三元组,n 个 agent 就需要 $n\times|S|^n\times|A|^n$ 的存储空间,可见学习过程的时间复杂度和空间复杂度都是 agent 个数的指数函数,所以一般 agent 个数不能太多,被求解问题的规模限制制约了此算法的应用范围.

RLG 方法的主要缺点是由于状态空间和动作空间的规模都是 agent 个数的指数函数,所以需要花费大量学习时间.为了加快 RLG 方法的收敛速度,研究人员进行了大量工作,并获得了一定成果[32].MBCL(memory based co-learning)和 TBCL(tree based co-learning)两种算法采用了结构化的存储方法,所以学习速度比一般的 RLG 方法有所改善[33].这两种算法相比,MBCL 的学习速度比 TBCL 快,而 TBCL 的存储空间耗费比 MBCL 要小.Nagayuki 等人的方法是 agent 先预测在某状态下另一个 agent 将执行什么动作,再决定自己应该相应地执行什么动作[34].

国内学者也对 RLG 方法进行了研究,南京大学的高阳等人提出了以非零和 Markov 游戏为框架的元对策 Q 算法,并证明了其收敛性,此算法仍然采用了组合状态和组合动作,所以也是一种 RLG 方法[35].

3 分布式强化学习方法中存在的问题及解决之道(Existing problems and their solution ways in distributed reinforcement learning algorithms)

1) 学习速度慢.

强化学习过程需要花费大量计算时间,分布式强化学习的计算量比标准强化学习的计算量还要大得多,尤其是 RLG 方法的学习空间是 agent 个数的指数函数,所以对强化学习的加速算法有着更迫切的需求.梯度法[36,37]、嵌入先验信息[38,39]、模糊学习[40]、状态空间划分[41~47]、分层学习[41,48,49]等现有的强化学习加速算法对特定任务背景依赖性很强,而且都是针对标准强化学习的,缺乏针对各类分布式强化学习特点的专用加速算法,更没有针对 RLG 方法的加速算法,所以还需要寻找更有效的分布式强化学习加速算法.

2) 连续状态空间及连续动作空间中的强化学习问题.

通常研究的强化学习系统,其状态和动作都认为是有限的集合,而实际问题中,其状态和动作往往是连续的.连续空间表示困难和学习速度慢的问题一直是实际应用中的巨大障碍,对连续空间的离散化有可能极大地损害学习的结果.文献[50~53]等分别提出了一些用于连续空间中强化学习的办法,但此问题一直没有较好的解决之道.

3) 结构信度分配问题.

信度分配问题包括时间信度分配和结构信度分配两方面,分布式强化学习主要考虑的是结构信度分配问题.结构信度分配的任务是将环境反馈回来的强化信号按照某种指标分配给系统中的所有 agent,使得性能高的 agent 具有更强的活性,性能低的 agent 逐渐被淘汰.桶群算法[54]是最常用的结构信度分配算法,另外还有其他一些针对模块化学习系统的结构信度分配算法[55,56]和针对分类器系统的结构信度分配算法等[57].这些算法在一定程度上解决了结构信度分配问题,但是这些算法都不是根据 agent 对任务的贡献分配强化值,而是根据它们以往收到的强化值折扣或其他类似指标进行分配,这种方法显然是不公平的,容易使得偶然体现出较高性能的 agent 不断获得较大的强化值份额,使得对它的性能评价不断循环增长.看来比较合理的结构信度分配方法是先将任务分解为子任务,再根据各个 agent 完成的子任务的重要程度给予奖励.另外目前的结构信度分配算法主要用于分类器系统和竞争学习系统,对协作学习系统支持还很不够.

4) RLI 与 RLG 的结合问题.

RLI 方法总的状态和动作空间是各个 agent 的状态和动作空间的简单相加,所以可以较容易地实现多个 agent 的 RLI 系统,而且此系统还可以动态增加或减少 agent 的个数,但是不易于实现协作或达到全局目标;RLG 方法形式简单,易于实现协作和达到全局目标,但是其学习空间是 agent 个数的指数函数,只能适用于 agent 很少的情况,而且不能动态

增加或减少 agent.由于这两种方法互有长短,却又相辅相成,所以如何取长补短,结合两者的优点就成为了重要课题,目前还没有这方面的研究成果,亟待开展研究工作.

4 总结与未来工作(Conclusion and future works)

本文分析比较了 RLI,RLG,RLS 三类分布式强化学习的研究成果、特点和适用范围,这三类方法各有各的适用范围,不能简单地分出优劣.目前这三类分布式强化学习方法的研究成果基本上都基于 Q 学习算法,几乎没有其他强化学习算法的分布式版本.

对目前存在的分布式强化学习方法,笔者认为:

1) RLI 是多个强化学习 agent 组成的群落,通过通信机制和协作机制将它们联合起来,有效的通信机制、协作机制和结构信度分配方案是 RLI 研究的主要方向.

2) RLG 采用组合状态和组合动作,注重全局利益,学习对整个 agent 团队最优的策略,但不利之处是 RLG 只适用于 agent 很少的情况,而且不能动态增加或减少 agent.如何消除组合爆炸,加速收敛是 RLG 研究的主要方向.

3) RLS 是 RLI 的推广,但 RLS 的通信机制、协作机制和结构信度分配方案是按照社会学和经济学的思想制定的,可以引入联盟、分组、选举、信赖、帮助、不耐烦、责备、夸奖等概念,体现出更高的 agent 智能和群体智能.RLS 系统通常比较复杂,适用于复杂的任务背景.RLS 研究的主要方向是继续模拟人类社会关系和经济运作规律,加大引入社会学和经济学思想的深度和广度,实现具有高智能的学习系统.

强化学习方法本质上是一种反应式的学习方法,很难理解事物之间的因果联系,进而分析做成某件事需要哪些先决条件,也无法归纳出一般的事物发展规律,从而把任务分解成子任务.分布式强化学习系统若要弥补这方面的不足,具备分析、归纳、推测、任务分解等方面的能力,就必须与归纳学习、类比学习等其他机器学习方法相结合,才能更好地理解多个 agent 之间的关系,使得其协作更加紧密,结构信度分配更加合理,从而达到更高的智能层次.可以预见,随着强化学习理论和多智能体理论的发展,分布式强化学习必将在更广阔的空间里获得更广泛的应用.

参考文献(References):

[1] SUTTON R. Learning to predict by the methods of temporal difference [J]. *Machine Learning*, 1988, 3(1): 9 - 44.

[2] WATKINS J, DAYAN P. Q-learning [J]. *Machine Learning*, 1992, 8(3/4): 279 - 292.

[3] 张汝波, 顾国昌, 刘照德, 等. 强化学习理论、算法及应用[J]. 控制理论与应用, 2000, 17(5): 637 - 642.
(ZHANG Rubo, GU Guochang, LIU Zhaode, et al. Reinforcement learning theory, algorithms and its application [J]. *Control Theory & Applications*, 2000, 17(5): 637 - 642.)

[4] CRITES R, BARTO A. Elevator group control using multiple reinforcement learning agents [J]. *Machine Learning*, 1998, 33(2/3): 235 - 262.

[5] KIM G, LEE G. Genetic reinforcement learning approach to the heterogeneous machine scheduling problem [J]. *IEEE Trans on Robotics and Automation*, 1998, 14(6): 879 - 893.

[6] MATARIC M. Using communication to reduce locality in multi-robot learning [A]. *Proc of the 14th National Conf on Artificial Intelligence* [C]. Menlo Park, CA: AAAI Press, 1997: 643 - 648.

[7] ARAI Y, FUJII T, ASAMA H. Adaptive behavior acquisition of collision avoidance among multiple autonomous mobile robots [A]. *Proc of IEEE/RSJ Int Conf on Intelligent Robots and Systems* [C]. Piscataway, NJ: IEEE Press, 1997: 1762 - 1767.

[8] KAWAKAMI K, OHKURA K, UEDA K. Adaptive role development in a homogeneous connected robot group [A]. *Proc of IEEE Int Conf on Systems, Man and Cybernetics* [C]. Piscataway, NJ: IEEE Press, 1999: 251 - 256.

[9] KOSTIADIS K, HU H. Reinforcement learning and cooperation in a simulated multi-agent system [A]. *Proc of IEEE/RSJ Int Conf on Intelligent Robots and Systems* [C]. Piscataway, NJ: IEEE Press, 1999: 990 - 995.

[10] SEN S, SEKARAN M, HALE J. Learning to coordinate without sharing information [A]. *Proc of the 11th National Conf on Artificial Intelligence* [C]. Menlo Park, CA: AAAI Press, 1994: 426 - 431.

[11] TOUZET C F. Distributed lazy Q-learning for cooperative mobile robots [J]. *Autonomous Robots*, 1999, 7(1): 1 - 14.

[12] TAN M. Multiagent reinforcement learning: independent vs cooperative agents [A]. *Proc of the 10th Int Conf on Machine Learning* [C]. [s.l.]: [s.n.], 1993: 330 - 337.

[13] NAKAYAMA T, MIKAMI S, WADA M. A realization of socially adaptive robots by competitive reinforcement learning [A]. *Proc of IEEE Int Workshop on Robot and Human Communication* [C]. Piscataway, NJ: IEEE Press, 1996: 107 - 111.

[14] BERENJI H R, SARAF S K. Competition and collaboration among fuzzy reinforcement learning agents [A]. *Proc of IEEE Int Conf on Fuzzy Systems* [C]. Piscataway, NJ: IEEE Press, 1998: 622 - 627.

[15] TAKADAMA K, NAKASUKA S, TERANO T. Multiagent reinforcement learning with organizational-learning oriented classifier system [A]. *Proc of IEEE Conf on Evolutionary Computation* [C]. Piscataway, NJ: IEEE Press, 1998: 63 - 68.

[16] MARIANO C E, MORALES E F. Distributed reinforcement learning for multiple objective optimization problems [A]. *Proc of IEEE Congress on Evolutionary Computation* [C]. Piscataway, NJ: IEEE Press, 2000: 188 - 195.

[17] HUNT J E, COOKE D E. An adaptive distributed learning system based on the immune system [A]. *Proc of IEEE Int Conf on Systems, Man and Cybernetics* [C]. Piscataway, NJ: IEEE Press, 1995: 2494 - 2499.

[18] 蔡庆生, 张波. 一种基于 agent 团队的强化学习模型与应用研究[J]. 计算机研究与发展, 2000, 37(9): 1087 - 1093.
(CAI Qingsheng, ZHANG Bo. An agent team based reinforcement learning model and its application [J]. *J of Computer Research and Development*, 2000, 37(9): 1087 - 1093.)

[19] KELLY I D, KEATING D A. Increased learning rates through the

sharing of experiences of multiple autonomous mobile robot agents [A]. *Proc of IEEE Int Conf on Fuzzy Systems* [C]. Piscataway, NJ: IEEE Press, 1998: 129 - 134.

[20] YAMAGUCHI T, MIURA M, YACHIDA M. Multiagent reinforcement learning with adaptive mimetism [A]. *Proc of IEEE Symposium on Emerging Technologies and Factory Automation* [C]. Piscataway, NJ: IEEE Press, 1996: 288 - 294.

[21] PRICE B, BOUTILIER C. Implicit imitation in multiagent reinforcement learning [A]. *Proc of the 16th Int Conf on Machine Learning* [C]. [s.l.]: [s.n.], 1999: 325 - 334.

[22] BUI H H, KIERONSKA D, VENKATESH S. Learning other agents' preference in multiagent negotiation [A]. *Proc of the 13th National Conf on Artificial Intelligence* [C]. Menlo Park, CA: AAAI Press, 1996: 114 - 119.

[23] SCHWARTZ R, KRAUS S. Negotiation on data allocation in multiagent environments [A]. *Proc of the 16th National Conf on Artificial Intelligence* [C]. Menlo Park, CA: AAAI Press, 1999: 29 - 35.

[24] CHEN Y, FU L, CHEN Y. Multi-agent based dynamic scheduling for a flexible assembly system [A]. *Proc of IEEE Int Conf on Robotics and Automation* [C]. Piscataway, NJ: IEEE Press, 1998: 2122 - 2127.

[25] FITOUSSI D, TENNENHOLTZ M. Choosing social law for multiagent systems: minimality and simplicity [J]. *Artificial Intelligence*, 2000, 119(1): 61 - 101.

[26] SHIBATA T, OHKAWA K, TANIE K. Spontaneous coordinated behavior of robots through reinforcement learning [A]. *Proc of IEEE Int Conf on Neural Networks* [C]. Piscataway, NJ: IEEE Press, 1995: 2908 - 2911.

[27] OHKAWA K, SHIBATA T, TANIE K. Self-generating method of behavioral evaluation for reinforcement learning among multiple coordinated robots[A]. *Proc of IEEE Int Conf on Intelligent Robots and Systems* [C]. Piscataway, NJ: IEEE Press, 1997: 1451 - 1456.

[28] SANDHOLM T. Approaches to winner determination in combinatorial auctions [J]. *Decision Support Systems*, 2000, 28(1/2): 165 - 176.

[29] BAUM E B. Toward a model of intelligence as an economy of agents[J]. *Machine Learning*, 1999, 35(2): 155 - 185.

[30] VIDAL J M, DURFEE E H. Learning nested agent models in an information economy [J]. *J of Experimental and Theoretical Artificial Intelligence*, 1998, 10(3): 291 - 308.

[31] HU J, WELLMAN M P. Multiagent reinforcement learning: theoretical framework and an algorithm [A]. *Proc of the 15th Int Conf on Machine Learning*[C]. [s.l.]: [s.n.], 1998: 115 - 122.

[32] CLAUS C, BOUTILIER C. The dynamics of reinforcement learning in cooperative multiagent systems [A]. *Proc of the 15th National Conf on Artificial Intelligence* [C]. Menlo Park, CA: AAAI Press, 1998: 746 - 752.

[33] SHEPPARD J W. Colearning in differential games [J]. *Machine Learning*, 1998, 33(2/3): 201 - 233.

[34] NAGAYUKI Y, ISHII S, DOYA K. Multiagent reinforcement learning: an approach based on the other agent's internal model [A]. *Proc of the 4th IEEE Int Conf on Multiagent Systems* [C]. Piscataway, NJ: IEEE Press, 2000: 215 - 221.

[35] 高阳, 周志华, 何桂洲, 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 Guizhou, et al. Research on Markov game-based multiagent reinforcement learning model and algorithm [J]. *J of Computer Research and Development*, 2000, 37(3): 257 - 263.)

[36] YAMAMURA M, ONOZUKA T. Reinforcement learning with knowledge by using a stochastic gradient method on a Bayesian network [A]. *Proc of IEEE Int Conf on Neural Networks* [C]. Piscataway, NJ: IEEE Press, 1998: 2045 - 2050.

[37] HO F, KAMEL M. Learning coordinated strategies for cooperative multiagent systems [J]. *Machine Learning*, 1998, 33(2/3): 155 - 177.

[38] HAILU G, SOMMER G. Embedding knowledge in reinforcement learning[A]. *Proc of the 8th Int Conf on Artificial Neural Networks* [C]. [s.l.]: [s.n.], 1998: 1133 - 1138.

[39] RIBEIRO C. Embedding a priori knowledge in reinforcement learning [J]. *J of Intelligent and Robotic Systems*, 1998, 21(1): 51 - 71.

[40] OH C, NAKASHIMA T, ISHIBUCHI H. Initialization of Q -values by fuzzy rules for accelerating Q -learning [A]. *Proc of IEEE Int Conf on Neural Networks* [C]. Piscataway, NJ: IEEE Press, 1998: 2051 - 2056.

[41] ISHIBUCHI H, NAKASHIMA T, MIYAMOTO H. Fuzzy Q -learning for a multi-player non-cooperative repeated game [A]. *Proc of IEEE Int Conf on Fuzzy Systems* [C]. Piscataway, NJ: IEEE Press, 1997: 1573 - 1579.

[42] SUN R, PETERSON T. Multiagent reinforcement learning: weighting and partitioning [J]. *Neural Networks*, 1999, 12(4): 727 - 753.

[43] TAKAHASHI Y, ASADA M, HOSODA K. Reasonable performance in less learning time by real robot based on incremental state space segmentation [A]. *Proc of IEEE/RSJ Int Conf on Intelligent Robots and Systems* [C]. Piscataway, NJ: IEEE Press, 1996: 1518 - 1524.

[44] HOUGEN D F, GINI M, SLAGLE J. Partitioning input space for reinforcement learning for control [A]. *Proc of IEEE Int Conf on Neural Networks* [C]. Piscataway, NJ: IEEE Press, 1997: 755 - 760.

[45] FINTON D J, HU Y. An application of importance-based feature extraction in reinforcement learning [A]. *Proc of the 4th IEEE Workshop on Neural Networks for Signal Processing* [C]. Piscataway, NJ: IEEE Press, 1994: 52 - 60.

[46] ASADA M, NODA S, HOSODA K. Action-based sensor space segmentation for soccer robot learning [J]. *Applied Artificial Intelligence*, 1998, 12(2/3): 149 - 164.

[47] SUN R, PETERSON T. Partitioning in reinforcement learning [A]. *Proc of Int Joint Conf on Neural Networks* [C]. Piscataway, NJ: IEEE Press, 1999: 1133 - 1138.

[48] ARAI Y, FUJII T, ASAMA H. Multilayered reinforcement learning for complicated collision avoidance problems [A]. *Proc of IEEE Int Conf on Robotics and Automation* [C]. Piscataway, NJ: IEEE Press,

1998:2186 - 2191.

[49] ENG K, ROBERTSON A, BLACKMAN D. Managing complexity in large learning robotic systems [J]. *J of Intelligence and Robotic Systems*, 2000,27(3):263 - 273.

[50] LEE J, OH S, CHOI D. TD based reinforcement learning using neural networks in control problems with continuous action space [A]. *Proc of the 8th Int Conf on Artificial Neural Networks* [C]. [s.l.]:[s.n.],1998:1133 - 1138.

[51] VOLLBRECHT H. Three principles of hierarchical task composition in reinforcement learning [A]. *Proc of the 8th Int Conf on Artificial Neural Networks* [C]. [s.l.]:[s.n.],1998:1121 - 1126.

[52] GROSS H, STEPHAN V, KRABBES M. A neural field approach to topological reinforcement learning in continuous action spaces [A]. *Proc of IEEE Int Conf on Neural Networks* [C]. [s.l.]: [s.n.],1998:1992 - 1997.

[53] ZANARDI C, HERVE J, COHEN P. A reactive planner for multiple robots involves in competitive tasks [A]. *Proc of IEEE Int Conf on Systems, Man and Cybernetics* [C]. Piscataway, NJ: IEEE Press, 1997:166 - 171.

[54] HOLLAND J H. Properties of the bucket brigade [A]. *Proc of 1st Int Conf on Genetic Algorithms* [C]. [s.l.]:[s.n.],1985:1 - 7.

[55] CHAPMAN K L, BAY J S. Task decomposition and dynamic policy merging in the distributed Q -learning classifier system [A]. *Proc of IEEE Int Symposium on Computational Intelligence in Robotics and Automation* [C]. Piscatawany, NJ: IEEE Press, 1997: 166 - 171.

[56] NORIHIKO O, OSAMU I, KENJI F. Acquisition of coordinated behavior by modular Q -learning agents [A]. *Proc of IEEE/RSJ Int Conf on Intelligent Robots and Systems* [C]. Piscatawany, NJ: IEEE Press,1996:1525 - 1529.

[57] BAY J, STANHOPE J. Distributed optimization of tactical actions by mobile intelligence agents [J]. *J of Robotic Systems*, 1997,14 (4):313 - 323.

作者简介:

仲 宇 (1976 —),男,分别于1998年、2001年在哈尔滨工程大学获学士和硕士学位,现为哈尔滨工程大学计算机科学与技术学院博士研究生.主要研究领域为计算智能,机器学习. E-mail:cleaner@0451.com;

顾国昌 (1946 —),男,1967年毕业于哈尔滨军事工程学院,1987年在法国获博士学位,现为哈尔滨工程大学计算机科学与技术学院教授,博士生导师.主要研究领域为智能机器人,机器人体系结构,行动决策和控制技术;

张汝波 (1963 —),男,分别于1984年、1987年和1999年在哈尔滨船舶工程学院、哈尔滨工程大学获学士、硕士和博士学位,现为哈尔滨工程大学计算机科学与技术学院副教授.主要研究领域为机器学习,计算智能,智能控制及智能机器人.

下 期 要 目