文章编号: 1001-0920(2011)11-1601-10

面向多机器人系统的增强学习研究进展综述

吴 军, 徐 昕, 王 健, 贺汉根

(国防科技大学 机电工程与自动化学院，长沙 410073)

摘 要: 基于增强学习的多机器人系统优化控制是近年来机器人学与分布式人工智能的前沿研究领域. 多机器人系统具有分布、异构和高维连续空间等特性, 使得面向多机器人系统的增强学习的研究面临着一系列挑战, 为此, 对其相关理论和算法的研究进展进行了系统综述. 首先, 阐述了多机器人增强学习的基本理论模型和优化目标; 然后, 在对已有学习算法进行对比分析的基础上, 重点探讨了多机器人增强学习理论与应用研究中的困难和求解思路, 给出了若干典型问题和应用实例; 最后, 对相关研究进行了总结和展望.

关键词: 多机器人系统；多智能体；增强学习；随机对策；马氏决策过程

中图分类号: TP24 **文献标识码:** A

Recent advances of reinforcement learning in multi-robot systems: A survey

WU Jun, *XU Xin*, *WANG Jian*, *HE Han-gen*

(College of Mechatronics and Automation，National University of Defense Technology，Changsha 410073，China. Correspondent：WU Jun，E-mail：aresnudt@yahoo.com.cn)

Abstract: Multi-robot optimization control based on reinforcement learning is a research frontier of robotics and distributed artificial intelligence in recent years. Some characteristics in multi-robot systems, such as distribution, heterogeneity and high-dimensional continuity, lead to a series of challenges in theoretical and methodological research for multi-robot reinforcement learning. Therefore, recent advances of multi-robot reinforcement learning are systematically surveyed. Firstly, the fundamental theoretical models and optimization objectives are analyzed. Based on a contrastive analysis for existing algorithms, the difficulties in theoretical research and implementations are discussed, and the possible solutions are summarized in detail. Several benchmark problems and applications are listed. Finally, current work and future research directions are concluded.

Key words: multi-robot systems; multi-agent; reinforcement learning; stochastic game; Markov decision process

1 引 言

多机器人系统(MRSs)在时空、功能、信息和资源上具有分布特性, 在任务适用性、经济性、最优性、鲁棒性和可扩展性等方面表现出极大的优越性, 从而在军事装备、工业生产和交通控制等领域具有良好的应用前景. 但是, 要使多机器人系统真正发挥其优势, 必须辅以合理的协同控制策略. 由于系统所处的实际工作环境往往非常复杂且难以建模, 环境知识获取困难, 导致基于模型和专家知识的传统方法难以取得好的控制效果. 因此, 研究具备自适应环境变化的学习控制方法对于多机器人系统极为重要[1].

增强学习(RL)是一种不依赖于环境模型和先验知识的机器学习方法, 通过试错和延时回报机制, 结合自适应动态规划方法, 能够不断优化控制策略, 为系统自适应外界环境变化提供了可行方案[2]. 通过将系统建模成马氏决策过程, 增强学习方法已成功地实现了单个机器人的优化控制[3-4]. 但是, 将增强学习方法推广应用于多机器人系统时, 由于多个学习器共存, 破坏了环境的平稳特性, 最终导致单智能体增强学习(SARL)的收敛性条件失效. 由于多机器人系统可抽象为多智能体系统(MASs), 多机器人增强学习(MRRL)可作为一类面向多机器人系统实际应用的多

收稿日期: 2011-03-13；**修回日期:** 2011-05-21.

基金项目: 国家自然科学基金项目(90820302, 61075072)；霍英东教育基金项目(114005)；湖南省自然科学基金项目(2007JJ3122).

作者简介: 吴军(1980−), 男, 博士生, 从事机器学习与智能系统的研究；徐昕(1974−), 男, 教授, 博士生导师, 从事机器学习、数据挖掘和机器人控制等研究.

智能体增强学习 (MARL) 展开研究. SARL, MARL 和 MRRL 三者的关系如图 1 所示. 显然, 三者具有一定的继承关系, 但学习复杂度依次递增, 相关理论及算法研究的成熟度依次递减. 如 SARL 中研究的学习算法可推广应用到后续问题中, 但会导致环境的非平稳性问题. MARL 通过引入随机对策模型可实现多个学习器之间的协同学习, 但加剧了维数灾难问题, 并导致结构信度分配、均衡点选择等新问题. 而 MRRL 需要处理实际物理系统引入的实时性、高度动态性和不确定性等问题, 学习难度更大, 最终导致相应研究工作主要停留于仿真环境或实验室结构化环境的算法验证上, 鲜见成功的实际应用案例.

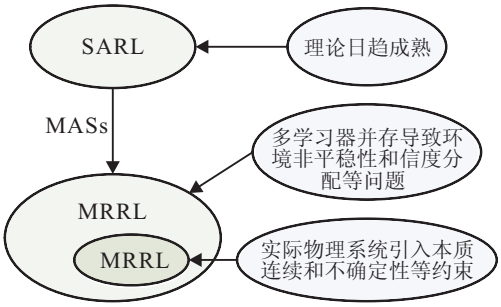

图 1 SARL, MARL 和 MRRL 之间的关系

将 MRRL 作为一类面向多机器人系统应用的 MARL, 并考虑实际物理系统引入的约束, 是当前 MRRL 的研究主线. 为了推动复杂和未知环境中的多机器人系统增强学习研究, 真正赋予多机器人系统自适应学习控制能力, 本文基于上述研究主线, 对近年来相关研究的主要进展和典型成果进行了总结和介绍, 并分析了相关挑战和应对措施, 可为科研人员提供参考.

2 多机器人增强学习的理论框架

2.1 模型框架基础

开展 MRRL研究需要合理的数学模型框架. 当前应用于 MRRL的模型框架可分为两大类: 应用于独立增强学习的马氏决策过程 (MDPs) 模型和应用于协同增强学习的随机对策 (SGs) 模型.

2.1.1 MDPs 模型

目前, 关于多机器人增强学习的研究中有一部分是直接应用 SARL 方法, 并获得了良好的优化控制效果[5], 而 MDPs 是该类学习方法的数学模型基础. 如图 2 所示, 一个 MDP 可定义为五元组 (S, A, P, R, γ). 其中: S 和 A 分别为状态集和行为集; P 为状态转移模型, $P(s, a, s')$ 表示在状态 s 处执行动作 a, 转移到新的状态 s' 的概率; $R: S\times A\to R$ 为回报函数; $\gamma\in[0,1)$ 为折扣因子.

增强学习的目的是最大化无限时间折扣总回报

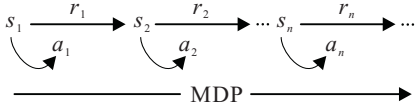

图 2 马氏决策过程模型

的期望值, 并获得对应的最优策略 π*. 为了对控制策略 π 进行评价, 可定义行为值函数 Q 如下:

$$Q^{\pi}(s,a)=E^{\pi}\Big(\sum_{t=0}^{\infty}\gamma^{t}r_{t}|s_{0}=s,a_{0}=a\Big). \tag{1}$$

对于任一 MDP, 可通过如下 Bellman 最优方程:

$$Q^{\pi^{*}}(s,a)=E[R(s,a)+\gamma\max_{a'}Q^{\pi^{*}}(s',a')], \tag{2}$$

确定最优策略[2]

$$\pi^{*}=\arg\max_{\pi}\ Q(s,\pi(s)). \tag{3}$$

2.1.2 SGs 模型

基于 MDPs 的独立增强学习方法没有考虑其他机器人对环境的影响, 在协同要求较高的任务中无法获得稳定的最优解. 因此, 当前 MRRL 正逐步转向基于 SGs 模型的协同增强学习的研究.

SGs 模型[6]中一个重要概念是矩阵对策, 可定义为多元组

$$\mathrm{MG}=\langle N, A_1, \cdots, A_N, R_1, \cdots, R_N\rangle.$$

其中: N 为玩家 (player) 数量, A_i 和 R_i $(i=1,2,\cdots,N)$ 分别为第 i 个玩家的有限行为集和回报函数. 每个玩家选择自己的行为 a_i, 但回报值 r_i 却由联合行为 $(a_1,\cdots,a_N)$ 决定. SGs 可由一个多元组 $\langle S, \mathrm{MG}, T, \gamma\rangle$ 描述. 其中: S 为有限联合状态集; MG 为 N 个玩家参与的矩阵对策; $T: S\times A_1\times\cdots\times A_N\to \Pi(S)$ 为状态转移函数, $\Pi(S)$ 为在状态空间 S 中的概率分布; $\gamma\in[0,1)$ 为折扣因子. 如图 3 所示, SGs 可看作 MDPs 与矩阵对策 MG 的结合, 是矩阵对策概念在多状态下的延伸.

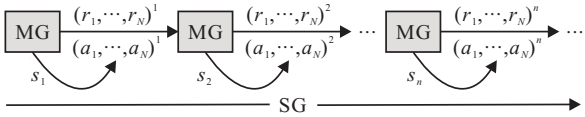

图 3 随机对策模型

在基于 SGs 模型的增强学习中, 每个智能体同样需要最大化自身折扣总回报的期望值, 但其下一状态和回报值由联合行为决定. 以包含 2 个智能体的增强学习为例, 智能体的平稳策略为 $\pi^i: S\times A^i\to[0,1], i=1,2$, 则给定初始状态 s 和联合策略 (π^1,π^2), 可定义第 i 个智能体的值函数为

$$V^{i}(s,\pi^{1},\pi^{2})=$$
$$E\Big(\sum_{t=0}^{\infty}\gamma^{t}r_{t}^{i}|s_{0}=s,\pi^{1},\pi^{2}\Big),\ i=1,2. \tag{4}$$

对应的行为值函数 Q 为

$$Q^i_{\pi^1,\pi^2}(s,a^1,a^2)=$$
$$r^i(s,a^1,a^2)+\gamma\sum_{s'\in S}T(s,a^1,a^2,s')V^i(s',\pi^1,\pi^2). \quad (5)$$

2.2 学习任务类型

根据任务类型的差异, 可将MARL任务分为静态任务和动态任务. 其中: 静态任务的状态不会发生转移, 并且可进一步分为无状态的对策学习任务和固定状态下的阶段对策学习任务; 动态任务学习过程中状态可发生转移, 主要针对多状态SGs问题进行学习. 当状态固定在某个特定状态时, 动态任务便转变成阶段对策学习任务.

重复对策学习与单次对策学习的主要差别在于: 针对静态任务, 重复对策学习通过多次迭代学习来获取其他智能体和回报函数的信息, 而单次对策学习则仅仅学习一次.

针对静态任务学到的是静态策略, 获取的是针对单状态或无状态对策问题的组合动作优化解. 而针对动态任务所学到的是序贯策略, 获取的是针对多状态SG问题的序贯动作优化解.

2.3 理论和方法基础

MARL综合采用了时间差分学习理论、对策论和直接策略搜索理论等理论工具. 图4描述了各种MARL算法的理论基础及相互关系[7].

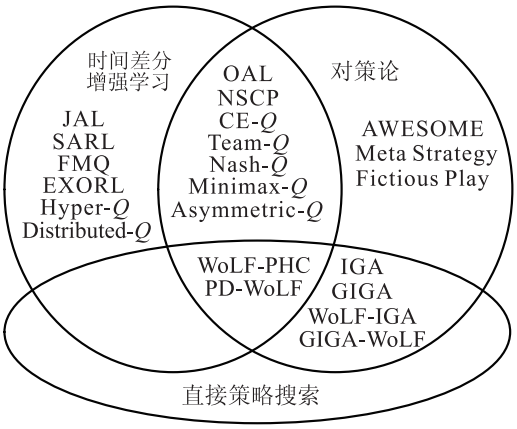

图 4　采用不同理论的MARL算法及其相互关系[7]

通过采用对策论, 可将动态任务分解为阶段对策问题, 并获得对应的阶段策略和期望回报, 但这样求解的合理性仍然存在争议[8-9]. 文献[10]采用元对策理论进行研究和分析, 使得智能体能根据自身愿望和对手策略估计进行策略修正, 既可以避免求解复杂的Nash均衡解, 又能得到全局最优解. 由于基于策略梯度的MARL方法研究较少[11], 本文将主要讨论基于值函数的增强学习技术.

2.4 均衡解概念

Nash均衡解对于MARL非常重要[12]. 针对N个玩家参与的对策问题, 其联合策略为$\pi=(\pi^1,\pi^2,\cdots,\pi^N)$, 如果任意玩家$i$都无法单独通过改变自身策略$\pi^i$来获得更好的回报$R^i$, 即

$$R^i(\pi)\geqslant R^i(\pi^1,\cdots,\pi^{i-1},\tilde{\pi}^i,\pi^{i+1},\cdots,\pi^N),\ \forall\tilde{\pi}^i, \quad (6)$$

则该联合策略π称为Nash均衡解. 任何静态对策问题至少存在一个Nash均衡解. 而且, 即使Nash均衡解是次优的, 收敛到Nash均衡解仍然是所有玩家的理性选择.

除最常用的Nash均衡解外, 还存在其他的均衡解概念, 如Correlated均衡解[13]和Stackelberg均衡解[14], 并且对应地可以构造CE-Q[13]和Asymmetric-Q学习方法[14]. 而基于Pareto最优解可得到Pareto-Q学习方法[15].

2.5 学习目标

学习目标对于MARL研究非常重要, 它支配了整个学习过程, 并为算法性能的评测和比较提供了依据. 由于理解上的差异, MARL学习目标定义各异, 但大致可分为稳定性目标[12]和适应性目标[16]两大类[7]. 稳定性本质上是指学习过程收敛到一个平稳策略的能力, 而适应性则指当前学习器适应其他学习器的策略改变, 并保持或提高自身学习性能的能力.下面将常见的MARL学习目标进行了综合和归类, 具体见表1.

表 1　学习目标的归类和统一

学习目标归类	各种学习目标定义
稳定性	收敛性[17-18]
	对手无关性[6]
	均衡解学习[19]
	预测[20]
适应性	理性[17]
	不遗憾性[18]
	对手理解[6]
	最佳响应学习[19]
	指定最优性/兼容性/安全性[16]

收敛到协同均衡解或平稳策略是一个基本的稳定性需求. 对手无关性以及均衡解学习都表示学习并稳定收敛到均衡解, 而预测则表示学习其他智能体的近似模型, 以提高学习的稳定性.

理性表示在给定其他智能体模型的条件下, 最大化期望回报, 或在其他智能体保持平稳策略时, 收敛到最佳响应策略. 不遗憾性要求智能体的回报值不差于任选一个平稳策略时获得的回报值, 从而可以防止被其他智能体欺骗和利用. 对手理解则是根据学到的其他智能体模型, 作出最佳响应学习. 指定最优性/兼容性/安全性表示在其他智能体分别采用固定策略和自身相同学习算法以及任意学习算法3种不同条件下, 当前智能体的学习性能均应满足相应级别的适应性要求.

总之, 对于性能良好的多机器人增强学习过程: 1) 稳定性目标是必需的, 且应兼顾渐近稳定性和暂态稳定性要求. 良好的稳定性有助于保证算法性能和进行算法分析, 并降低其他学习器所处环境的非平稳性. 2) 适应性目标也是必须的. 良好的适应性可防止被其他学习器恶意欺骗和利用. 3) 完美的稳定性和适应性不可兼得. 一般在学习过程中满足预设的学习目标边界条件即可.

3 多机器人增强学习方法的分类研究

3.1 多机器人增强学习的分类

多机器人增强学习方法的分类标准包括: 所采用学习算法的异同, 基于模型与否或者协同关系等. 本文根据模型框架的不同, 分为基于MDPs模型和基于SGs模型的MRRL两大类. 前者又分为集中式和分布式独立MRRL两类; 后者根据协作关系的异同, 可进一步分为3类: 基于共同回报随机对策(CISG)模型[21], 基于零和随机对策(ZSSG)模型以及基于一般和随机对策(GSSG)模型[6]的增强学习方法. 具体分类如图5所示.

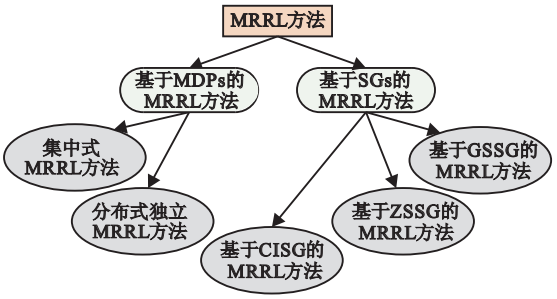

图 5 多机器人增强学习方法分类示意图

CISG模型主要针对全合作的学习任务, 在此种学习任务中, 所有学习器的回报值都相等, 即 $r_1=r_2=\cdots=r_N$; ZSSG模型主要针对对抗性任务, 所有学习器的回报值之和恒为零, 即 $\sum_{i=1}^{N} r_i=0$; 而GSSG模型则用于一般的协作行为, 对回报值没有附加额外的约束条件.

3.2 各类方法研究现状

3.2.1 集中式多机器人增强学习方法

该学习方法适用于采用分散感知、集中学习组织形式的多机器人系统, 由一个中央学习器负责学习, 其他机器人不是学习主体, 仅将感知的环境状态提交给中央学习器, 并接收和执行动作命令. 如联合状态 $s=(s^1,\cdots,s^N)$, 联合行为 $a=(a^1,\cdots,a^N)$, 中央学习器采用 Q 学习规则, 时间步为 k, 则有

$$Q_{k+1}(s_k,a_k)=$$
$$Q_k(s_k,a_k)+$$
$$\alpha[r_{k+1}+\gamma\max_{a'}Q_k(s_{k+1},a')-Q_k(s_k,a_k)], \quad (7)$$

其动作策略为

$$\pi(s_k,\cdot)=\arg\max_a Q_k(s_k,a). \quad (8)$$

其他机器人 i 从中央学习器决策的联合动作 a 获取对应的子动作 a_k^i, 并加以执行即可.

集中式增强学习方法常应用于自身处理能力有限的多机器人系统, 必须借助中央服务器实现学习控制, 如小型组的足球机器人系统[22-23]. 集中式学习方法随系统规模的增加, 将面临学习空间的规模爆炸问题. 此外, 学习集中化与数据内在分布性之间的矛盾也导致其对通信的依赖度很高.

3.2.2 分布式独立多机器人增强学习方法

该学习方法中的机器人直接将其他机器人视为环境的一部分, 并采用标准SARL方法进行学习. 分布式独立学习方法简单易用, 不存在信度分配问题, 可扩展性好, 适用于大规模的集群多机器人系统, 其典型应用包括觅食[1]、巡逻[24-25]等.

分布式独立学习器采用 Q 学习规则, 各自独立感知和决策, 则第 i 个学习器的算法描述为

$$Q_{k+1}^i(s_k^i,a_k^i)=$$
$$Q_k(s_k^i,a_k^i)+$$
$$\alpha[r_{k+1}^i+\gamma\max_{a'\in A^i}Q_k(s_{k+1}^i,a')-Q_k(s_k^i,a_k^i)], \quad (9)$$

$$\pi^i(s_k^i,\cdot)=\arg\max_{a\in A^i}Q_k^i(s_k^i,a), \quad (10)$$

其中 s_k^i 和 a_k^i 分别为第 i 个学习器的状态和动作. 采用该学习方法的机器人以自我为中心, 没有考虑其他机器人对环境的影响, 收敛性难以保证, 不易获得全局最优解, 仅适用于弱协同的多机器人任务系统. 关于分布独立MRRL方法的理论探讨可参见文献[26].

3.2.3 基于CISG的多机器人增强学习方法

该方法采用全局回报机制均分增强信号, 适用于强协作紧耦合的多机器人任务. 基于CISG的协同学习方法[21]有Team-Q[6], Distributed- Q[27]和OAL[28]等. 由于回报值均相同, 即 $r_1=r_2=\cdots=r_N$, 有

$$Q^1=Q^2=\cdots=Q^N. \quad (11)$$

如果存在集中式中央学习器, 则其值函数 Q 更新规则如式(7)和(8)所示, 但由于各个学习器可以独立学习和决策, 导致面临多均衡点协同问题. 如Team-Q算法[6]便无法处理多平衡点问题, 且算法复杂度和系统规模呈指数关系, 适用范围有限. Distributed-Q[27]算法采用乐观策略单调更新本地 Q 函数, 可收敛到最优联合策略, 但只能处理确定性问题. 最优自适应学习算法OAL[28]将多状态的团队对策分解为阶段对策, 并建立虚对策来消除所有严格次优纳什均衡解, 是唯一能够确保找到CISG问题最优解的学习算法, 但计算和存储代价高昂. JAL[29]算法和FMQ[20]算

法利用模型信息或者历史统计数据, 启发并引导学习器选择更好的协同动作, 但二者均只适用于静态任务的学习.

3.2.4 基于 ZSSG 的多机器人增强学习方法

该学习方法适合于对抗性的多机器人学习任务. 常用的 Minimax-Q 算法[6]采用 Minimax 原理来计算阶段对策问题的策略和值函数, 并采用时间差分学习规则实现行为函数值的学习.

以包含两个学习器的多机器人系统为例, 学习器 $i\,(i=1,2)$ 中的 Minimax-Q 算法可描述如下:

$$\begin{aligned}&Q^i_{k+1}(s_k,a_k)=\\&Q^i_k(s_k,a_k)+\\&\alpha[r^i_{k+1}+\gamma\cdot\mathrm{Eval}^i(Q^i_k,s_{k+1})-Q^i_k(s_k,a_k)],\end{aligned}\tag{12}$$

$$\pi^i_k(s_k,\cdot)=\mathrm{Solver}^i(Q^i_k,s_k).\tag{13}$$

式中 Eval^i 和 Solver^i 分别描述如下:

$$\begin{cases}\mathrm{Eval}^i(Q^i,s)=\\ \max\limits_{\pi^i(s,\cdot)}\min\limits_{a^{i-}}\sum\limits_{a^i}\pi^i(s,a^i)Q^i(s,a^i,a^{i-}),\\ \mathrm{Solver}^i(Q^i,s)=\arg\mathrm{Eval}(Q^i,s).\end{cases}\tag{14}$$

其中: $\pi^i_k(s_k,\cdot)$ 为 k 时刻学习器 i 在联合状态 s_k 处采取的随机策略, a^i_k 为学习器 i 在 k 时刻采取的具体动作. 因为 $r^1=-r^2$, 所以有

$$Q=Q^1=-Q^2.$$

Minimax-Q 算法假定对手是理性的, 从而导致学习过程过于保守, 最终解往往是无价值和不合理的, 如果引入对手模型将有可能获得更优解.

3.2.5 基于 GSSG 的多机器人增强学习方法

GSSG 模型可反映个体理性与集体理性冲突的本质特性, 更适用于描述自利型多智能体系统. 基于 GSSG 的 MRRL 方法适用于弱协作紧耦合的多机器人任务, 而且可以根据当前学习器与其他学习器的关系分为 3 类[7]: 相互无关学习方法、基于跟踪的学习方法以及基于理解的学习方法. 具体分类和特性如表 2 所示.

相互无关学习方法首先将 SG 问题分解为阶段对策问题, 再利用对策原理和 Q 学习方法求取各个学习器的值函数和策略, 即

$$\begin{aligned}&Q^i_{k+1}(s_k,a_k)=\\&Q^i_k(s_k,a_k)+\\&\alpha[r^i_{k+1}+\gamma\cdot\mathrm{Eval}^i(Q_k,s_{k+1})-Q^i_k(s_k,a_k)],\end{aligned}\tag{15}$$

$$\pi^i_k(s_k,\cdot)=\mathrm{Solver}^i(Q_k,s_k).\tag{16}$$

其中

$$\begin{cases}\mathrm{Eval}^i(Q_k,s_k)=V^i(s_k,\mathrm{EQ}(Q_k(s_k,\cdot))),\\ \mathrm{Solver}^i(Q_k,s_k)=\mathrm{EQ}^i(Q_k(s_k,\cdot)).\end{cases}\tag{17}$$

s_k,a_k 表示联合状态和联合动作, $\mathrm{EQ}(\cdot)$ 表示求解均衡解, $\mathrm{EQ}^i(\cdot)$ 和 Solver^i 均表示返回均衡解的第 i 个解分量, Eval^i 求解的是该均衡解分量的期望值. 均衡解求解算子 $\mathrm{EQ}(\cdot)$ 可分别为 Nash 均衡解、Correlated 均衡解和 Stackelberg 均衡解求解算子, 并对应构成 Nash-Q[12], CE-Q[13]和Asysmetric-Q[14]算法. 相互无关学习方法只是消极地收敛到均衡解, 并没有考虑其他 Agent 的策略, 不能保证解策略的最优性.

基于跟踪的学习方法是通过估计对手的策略模型, 并采取最佳响应动作. 对手策略模型可通过下式求取:

$$\hat{\pi}^i_j(s,a^j_k)=\frac{C^i_j(s,a^j_k)}{\sum\limits_{\hat{a}\in A^j}C^i_j(s,\hat{a})},\tag{18}$$

其中 $C^i_j(s,a^j_k)$ 表示第 i 个学习器观察第 j 个学习器在状态 s 执行动作 a^j_k 的次数. 具体地, 如 NSCP[31]算法, 其计算与式 (15) 和 (16) 类似, 差别在于

$$\begin{cases}\mathrm{Eval}^i(Q_k,s_k)=\mathrm{BR}^i(Q^i_k,s_k),\\ \mathrm{Solver}^i(Q_k,s_k)=\arg\mathrm{BR}^i(Q^i_k,s_k).\end{cases}\tag{19}$$

其中

$$\begin{aligned}&\mathrm{BR}^i(Q^i,s)=\\&\max_{\pi^i(s,\cdot)}\sum_a\Big[\pi^i(s,a^i)\prod_{j=1,j\neq i}^{N}\hat{\pi}^i_j(s,a^j)Q^i(s,a)\Big].\end{aligned}\tag{20}$$

基于理解的学习方法兼顾了收敛性和适应性. 如 AWESOME[35]算法, 实时监视对方策略, 若对方的策略为平稳策略, 则采用基于最佳响应的 Fictitious Play 算法[19]; 若对方策略不平稳, 则采用保守的 Nash 均衡解算法[19]. 此外, 基于理解的学习算法常采用直接策略搜索方法确保收敛性, 如 IGA[32], WoLF-IGA[19],

表 2　基于 GSSG 模型的多机器人增强学习方法分类

分　类	特　性	算法实例
相互无关学习方法	共享学习结构; 动作和回报完全可观测; 将 SGs 问题分解为阶段对策求解; 强调学习的收敛性	Nash-Q[12], CE-Q[13], Asymmetric-Q[14]
基于跟踪的学习方法	动作完全可观测; 需要估计对手模型, 相应作出最佳响应决策, 强调适应性	MetaStrategy[16], Hyper-Q[30], NSCP[31]
基于理解的学习方法	常结合梯度信息和启发式规则; 兼顾学习收敛性和适应性	IGA[32], GIGA[33], WoLF-IGA[19], GIGA-WoLF[18], WoLF-PHC[19], EXORL[34]

GIGA[33], GIGA-WoLF [18]和 WoLF-PHC[19]算法. 然而 EXORL[34]是一种基于 SARSA 的算法, 通过采用互补思想, 使得无论其他学习器采取自适应策略还是固定策略, 它都能够学习到最佳响应策略.

4 多机器人增强学习的难点及发展趋势

多机器人增强学习以 MARL为理论基础, 不但需面临 MARL 的各种固有难题, 而且需处理实际物理系统引入的各种问题.

4.1 多智能体增强学习的固有难题

在 MARL 中, 由于环境中存在多个学习器, 使得 MARL 问题变得复杂, 主要的有如下难题:

1) 维数灾难问题. 除分布式独立 MARL 方法和少数基于 SGs 模型的 MARL 算法 (如 Distributed-Q[27]和 Hysteretic-Q[36]等) 外, 基本上所有基于值函数的 MARL 算法都面临维数灾难问题, 即学习复杂度与智能体数量呈指数增长关系.

2) 信度分配问题. MARL 的信度分配问题包括时间信度分配和结构信度分配两方面[37]. 时间信度分配将最终的增强信号转化为随时间分布的增强信号序列, 以便对动作序列中的每个动作作出评价; 结构信度分配将增强信号标量转化为增强信号向量, 以对同时行动的多个学习系统分别作出评价. MARL 的信度分配难题包括:

① 稀疏化的增强信号导致学习缓慢, 开发合理的时间信度分配方法, 加快学习收敛成为一大难题. 基本上所有基于联合状态-动作的 MARL 学习算法都面临巨大学习空间与稀疏增强信号之间的矛盾.

② 简单的全局信度分配和局部信度分配方法均导致学习效果不佳, 开发合理的结构信度分配方案, 突破 SARL/MARL 之间的屏障成为一大难题. 而基于 CISG 的 MARL 学习方法采用信度平均分配机制, 无法对每个参与者的动作进行精确评价, 不利于学习的快速收敛.

3) 多均衡点协同选择问题. 基于 CISG/GSSG 模型的 MARL 算法, 如 Team-Q[6]和 Nash-Q[12], 大多存在多均衡解的问题, 必须有合理的均衡解协同选择机制, 才能保证最终解的稳定性和优化性.

4.2 物理系统引入的约束限制

1) 学习过程的危险性. 虽然增强学习方法基于试错原理, 而且容许机器人与外部环境进行交互学习时犯错, 但物理机器人系统可能因为鲁莽的随机探索行为而导致毁灭性的后果, 尤其在多机器人系统中, 不但环境可对机器人造成伤害, 机器人之间也可能造成严重伤害, 所以在多机器人增强学习研究中不可盲目采用试错法. 在保证充分探索时, 还应限定机器人的安全探索范围, 满足各种安全性约束, 确保机器人的安全.

2) 学习代价的高昂特性. 在 MARL 研究中, 更多地考虑理论上的收敛性和优化性, 而对智能体执行动作所花费的时间、能量等代价常常忽略不计. 但在多机器人增强学习任务中, 物理机器人执行任务的过程相对缓慢, 需要消耗大量的时间和资源, 而增强学习过程中动辄上万次的样本搜集与迭代学习过程, 将导致学习代价过于高昂. 所以, 开发合理的样本采集策略快速遍历学习空间, 同时开发能高效利用有限学习样本的快速学习算法已成为迫切需要.

3) 学习空间的连续性. MARL 的很多算法都是针对离散状态和离散动作设计的, 而在多机器人系统中, 状态和动作的内在连续特性对学习算法的开发应用是极大的考验.

4) 学习的在线特性和异步性. 机器人大多要求进行在线学习, 如何实现在线的迭代递增学习过程, 并满足稳定性和实时性要求极为重要. 此外, MARL 中大多假设在理想环境中可实现学习决策的严格同步, 但实际机器人的处理能力各异, 导致完成动作所需时间不一致, 学习同步操作比较困难.

5) 系统的高度动态性和不确定性. MASs 的环境可建模为理想化环境, 而 MRSs 的学习环境必定是一个动态系统, 这便对仅考虑学习器动力学特性的 MARL 方法提出了挑战[38]. 此外, 机器人感知和执行过程中存在各种干扰和噪声, 且面临部分可观测问题, 这都导致多机器人系统对学习过程的鲁棒性提出了更高的要求.

4.3 多机器人增强学习的发展趋势分析

虽然多机器人增强学习面临诸多困难, 但推进其实用化的研究工作仍在持续开展, 下面就相关研究的前沿性工作进行总结和论述.

1) 准确定位适用范围. 增强学习方法有适用的任务对象, 虽然增强学习算法不依赖于模型, 简单易用, 但依据 NFL 定理[39], 有得必有失, 如增强学习存在对系统规模敏感, 且学习速率慢等不足. 如将增强学习方法应用于高层或底层的系统控制任务 (如高层的角色分配优化和底层的具体控制任务优化), 则由于学习空间相对简单, 可以更显著地发挥增强学习的优势[40]. 因此, 必须摒弃滥用增强学习方法的思维方式.

2) 复杂学习问题的简化. 多机器人增强学习的学习复杂度主要源于其巨大的学习空间. 为了保证学习效果, 可对问题进行适度的简化操作, 如学习任务的分解、状态压缩、结合反应式方法以及采用结构化、分层化的学习方法等, 具体见表 3.

3) 提高学习效率. 受机器人物理运行过程的限

制, 学习样本搜集代价高昂, 学习过程极其缓慢, 所以应提高有限样本的利用效率, 并采用各种学习加速方法, 以保证增强学习优化控制效果的及早发挥 (见表 4).

4) 其他多机器人增强学习问题求解. 对于其他的多机器人增强学习研究的难点, 表 5 列出了相关应对及求解方法.

5) 注重与其他方法的结合. 将增强学习方法与其他优化方法结合应用也可实现优势互补, 提高问题求解性能. 如将遗传算法与增强学习算法相结合[70,78], 可以更有效地解决多机器人增强学习中的局部极小问题.

表 3　多机器人增强学习问题的简化方法

分　类	原　理	具体方法
任务分解	将复杂学习任务进行分解, 简化学习任务难度	直接分解法[41]; 协同图方法[42], Sparse-Q方法[43], 基于模块化仲裁的分解法[44]
状态压缩	采用信息归纳或状态分组, 消除冗余状态, 增大学习的空间粒度	有色轨道划分法和基于状态分量重要性划分法[37], 模糊化状态法[45-46], 自组织压缩映射法[47], 定性描述法[48]
结合反应式方法	结合基于行为的方法简化动作空间, 并保持底层控制的连续性	基于行为的 Q 学习方法[1,49-50], rQ 学习方法[51]
结构化及分层方法	利用“抽象”机制, 实现任务分割或引入宏动作, 并实现子策略的复用	结构化增强学习方法[52-53], 分层学习方法[54]

表 4　多机器人增强学习的学习加速方法

途　径	原　理	具体方法
模仿与知识共享	模仿可直接获得次优可行策略; 知识可避免重复学习, 减少学习代价	模仿学习[55-56], 基于案例推理的方法[57-58], 共享传感信息、经验序列、策略的方法[46] , 迁移学习[59-60]
启发式探索技术	利用学习得到的简化环境模型, 规划获得次优策略, 引导探索操作	启发式探索方法[61]
结合先验知识	利用相关领域知识, 简化学习问题复杂度	预编程的反射行为[1], Shaping技术[62], 任务结构分解[52]

表 5　其他多机器人增强学习问题的应对方法

学习难点	求解措施
连续学习空间	采用泛化逼近技术, 如神经网络逼近[63]、基函数线性逼近[64]、核函数逼近技术[65-66]
在线学习	将在线 LSPI 算法[67]和在线KLSPE算法[68] 拓展应用到多机器人增强学习中
学习同步与不确定性问题	引入 SMDP 模型[24]处理任意长时间步的学习过程, 利用POMDPs模型[69]处理不确定性问题
协同问题	基于习俗、角色或通信的直接协同方法, 如顺序 Q 方法[70]; 基于对手建模的间接协同方法, 如随机学习自动机[37], Bayesian 学习[71]和极大似然估计[72]方法
信度分配	对于时间信度分配问题, 采用平均回报方法[73], 过程奖赏函数方法[74], 基于中期目标以及局部目标的奖赏函数方法[75]; 对于结构信度分配问题, 采用轮流学习的方法, 回报滤波方法[76], 强化信号内外分解方法[77]

5　典型应用领域

多机器人增强学习研究的典型应用领域包括多机器人足球、协同搬运、多目标观测、搜索与覆盖以及追逃问题研究, 具体见表 6.

对于搜索和覆盖等协同性要求比较低的多机器人任务, 可采用分布式独立多机器人增强学习方法, 以减小学习空间, 加快学习速率, 但可能牺牲解的稳定性与最优性. 而对于推箱子和协同搬运等协同性要求很高的多机器人任务, 则多采用基于 CISG 的多机器人增强学习方法, 如 Team-Q, 以保证解的最优性, 但学习速率慢, 且多需要通信等协同机制[70]. 对于追逃和机器人足球, 可以在内部采用基于 CISG 的多机器人增强学习方法, 以实现合作, 而作为一个团体执行对抗任务时, 则采用基于 ZSSG 的多机器人增强学习的方法. 对于任务性质不明的多机器人任务, 可以考虑采用基于 GSSG 的多机器人增强学习方法, 比如 AWESOME[35]算法.

总体而言, 当前关于多机器人系统的增强学习研

表 6　典型的多机器人增强学习应用及实例

应　用	特　点	相关实例和文献
多机器人足球	多集中于高层的协调控制, 应用最广	行为、角色等对象的协调控制[23,46,54,79-80]
协同搬运	搬运物体, 多集中于底层的动作协同控制, 协同性要求高	推箱子[81], 协同搬运[70,78]
多目标观测	检测或跟踪一组移动目标, 并最大化正常监视的时间	巡逻[25], 多目标跟踪[50]
搜索与覆盖	主要实现协调优化控制, 协同性要求较低	觅食[1,49], 行星探索[82]
追逃	既可纯对抗, 也可对抗兼合作	追逃[83-84]

究大多集中于算法理论和仿真验证研究, 实际的实物验证也大多集中于实验室结构化环境, 所以推进室外非结构化环境中的应用研究是多机器人增强学习下一步研究的重点.

6 结论与思考

通过前述总结和分析可知, 当前多机器人增强学习研究呈现如下特点:

1) 多机器人增强学习研究已成为提高系统自适应控制能力的有效途径, 并将成为必然的研究方向.

2) 目前关于多机器人增强学习的研究主要在吸收MARL理论和方法的基础上再考虑解决实际物理系统引入的各类约束问题.

3) 当前的相关研究可分为基于MDPs模型和基于SGs模型的多机器人增强学习两大类. 实际应用中多采用分布式独立学习方法, 而协同学习方法需要结合问题简化和学习加速技术, 以提高学习效果.

4) 基于SGs模型的多机器人增强学习方法, 一般先将多状态的序贯学习问题分解为阶段对策学习问题, 再加以求解, 但其合理性仍然存在争议.

5) 当前的研究仍以仿真研究或实验室结构化环境中的简单验证研究为主, 实际应用亟待推进.

随着学习理论和相关技术的发展, 相信在不久的将来, 多机器人增强学习的理论发展和实用化水平将迈上一个新的台阶.

参考文献(References)

[1] Mataric M J. Reinforcement learning in the multi-robot domain[J]. Autonomous Robots, 1997, 4(1): 73-83.

[2] Szepesvari C. Algorithms for reinforcement learning: Synthesis lectures on artificial intelligence and machine learning[M]. San Rafael: Morgan & Claypool Publishers, 2009.

[3] 段勇, 杨淮清, 崔宝侠, 等. 强化学习在足球机器人基本动作学习中的应用[J]. 机器人, 2008, 30(5): 453-459.
(Duan Y, Yang H Q, Cui B X, et al. Application of reinforcement learning to basic action learning of soccer robot[J]. Robot, 2008, 30(5): 453-459.)

[4] 段勇, 崔宝侠, 徐心和. 进化强化学习及其在机器人路径跟踪中的应用[J]. 控制与决策, 2009, 24(4): 532-536.
(Duan Y, Cui B X, Xu X H. Evolutionary reinforcement learning and its application in robot path tracking[J]. Control and Decision, 2009, 24(4): 532-536.)

[5] Paternina-Arboleda C D, Das T K. A multi-agent reinforcement learning approach to obtaining dynamic control policies for stochastic lot scheduling problem[J]. Simulation Modeling Practice and Theory, 2005, 13(5): 389-406.

[6] Littman M L. Value-function reinforcement learning in Markov games[J]. J of Cognitive Systems Research, 2001, 2(1): 55-66.

[7] Busoniu L, Babuska R, de Schutter B. A comprehensive survey of multi-agent reinforcement learning[J]. IEEE Trans on Systems, Man, and Cybernetics — Part C: Applications and Reviews, 2008, 38(2): 156-172.

[8] Shoham Y, Powers R, Grenager T. Multi-agent reinforcement learning: A critical survey[R]. Stanford: Computer Science Department, 2003.

[9] Panait L, Luke S. Cooperative multi-agent learning: The state of the art[J]. Autonomous Agents and Multi-agent Systems, 2005, 11(3): 387-434.

[10] 王皓, 高阳. 元博弈平衡和多Agent强化学习的Meta Q算法[J]. 计算机研究与发展, 2006, 43(增): 137-141.
(Wang H, Gao Y. Meta-equilibria and Meta Q algorithm for multi-agent reinforcement learning[J]. J of Computer Research and Development, 2006, 43(S): 137-141.)

[11] Ma J, Cameron S. Combining policy search with planning in multi-agent cooperation[C]. RoboCup 2008: Robot Soccer World Cup XII. Berlin: Springer, 2009: 532-543.

[12] Hu J, Wellman M P. Nash Q-learning for general-sum stochastic games[J]. J of Machine Learning Research, 2003, 4(11): 1039-1069.

[13] Greenwald A, Hall K. Correlated-Q learning[C]. Proc of ICML-03. Washington, 2003: 242-249.

[14] Kononen V. Asymmetric multi-agent reinforcement learning[J]. Web Intelligence and Agent Systems, 2004, 2(2): 105-121.

[15] Song M, Gu G, Zhang G. Pareto-Q learning algorithm for cooperative agents in general-sum games[C]. Proc of CEEMAS2005, LNAI 3690. Berlin: Springer, 2005: 576-578.

[16] Powers R, Shoham Y. New criteria and a new algorithm for learning in multi-agent systems[C]. Proc of NIPS'04. Cambridge: MIT Press, 2005: 1089-1096.

[17] Bowling M H, Veloso M M. Multi-agent learning using a variable learning rate[J]. Artificial Intelligence, 2002, 136(2): 215-250.

[18] Bowling M. Convergence and no-regret in multiagent learning[C]. Proc of NIPS'04. Cambridge: MIT Press, 2005: 209-216.

[19] Bowling M. Multiagent learning in the presence of agents with limitations[R]. Pittsburgh: School of Computer Science, Carnegie Mellon University, 2003.

[20] Kapetanakis S, Kudenko D. Reinforcement learning of coordination in heterogeneous cooperative multi-agent systems[C]. Adaptive Agents and MAS II. Berlin: Springer-Verlag, 2005: 119-131.

[21] Bab A, Brafman R I. Multi-agent reinforcement learning in common interest and fixed sum stochastic games: An experimental study[J]. J of Machine Learning Research, 2008, 9(12): 2635-2675.

[22] 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 Q S, Zhang B. An agent team based reinforcement learning model and its application[J]. J of Computer Research and Development, 2000, 37(9): 1087-1093.)

[23] Sanz Y, de Lope J, Martin H J A. Applying reinforcement learning to multi-robot team coordination[C]. Proc of HAIS'08, LNCS 5271. Heidelberg: Springer-Verlag, 2008: 625-632.

[24] Sutton R, Precup D, Singh S. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning[J]. Articial Intelligence, 1999, 112(1/2): 181-211.

[25] Santana H, Ramalho G, Corruble V, et al. Multi-agent patrolling with reinforcement learning[C]. Proc of AAMAS'04. Washington: IEEE Press, 2004: 1122-1129.

[26] Tuyls K, Hoen P J' T, Vanschoenwinkel B. An evolutionary dynamical analysis of multi-agent learning in iterated games[J]. Autonomous Agents and Multi-agent Systems, 2006, 12(1): 115-153.

[27] Lauer M, Riedmiller M. An algorithm for distributed reinforcement learning in cooperative multi-agent systems[C]. Proc of ICML-00. San Francisco: Morgan Kaufmann, 2000: 535-542.

[28] Wang X, Sandholm T. Reinforcement learning to play an optimal Nash equilibrium in team Markov games[C]. Proc of NIPS'03. Vancouver and Whistler, 2004: 1571-1578.

[29] Claus C, Boutilier C. The dynamics of reinforcement learning in cooperative multiagent systems[C]. Proc of AAAI/IAAI-98. Madison, 1998: 746-752.

[30] Tesauro G. Extending Q-learning to general adaptive multi-agent systems[C]. Proc of NIPS'03. Vancouver and Whistler, 2004: 871-878.

[31] Weinberg M, Rosenschein J S. Best-response multi-agent learning in non-stationary environments[C]. Proc of AAMAS'04. New York, 2004: 506-513.

[32] Singh S, Kearns M, Mansour Y. Nash convergence of gradient dynamics in general-sum games[C]. Proc of UAI-00. San Francisco, 2000: 541-548.

[33] Zinkevich M. Online convex programming and generalized infinitesimal gradient ascent[C]. Proc of ICML-03. Washington, 2003: 928-936.

[34] Suematsu N, Hayashi A. A multiagent reinforcement learning algorithm using extended optimal response[C]. Proc of AAMAS'02. Bologna, 2002: 370-377.

[35] Conitzer V, Sandholm T. AWESOME: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents[C]. Proc of ICML-03. Washington, 2003: 83-90.

[36] Matignon L, Laurent G J, Fort-Piat N L. Hysteretic Q-learning: An algorithm for decentralized reinforcement learning in cooperative multi-agent teams[C]. Proc of IROS'07. San Diego, 2007: 64-69.

[37] 仲宇. 学习理论及在多机器人中的应用研究[D]. 哈尔滨: 哈尔滨工程大学计算机科学与技术学院, 2003.
(Zhong Y. Research on distributed reinforcement learning theory and its applications in multi-robot systems[D]. Harbin: College of Computer Science and Technology, Harbin Institute of Technology, 2003.)

[38] Tuyls K, Nowe A. Evolutionary game theory and multi-agent reinforcement learning[J]. The Knowledge Engineering Review, 2005, 20(1): 63-90.

[39] Wolpert D H, William G M. No free lunch theorems for optimization[J]. IEEE Trans on Evolutionary Computation, 1997, 1(1): 67-82.

[40] 赵杰, 姜健, 臧希喆. 基于强化学习的多机器人编队导航[J]. 辽宁工程技术大学学报, 2007, 26(6): 915-918.
(Zhao J, Jiang J, Zang X Z. Multi-robots formation and navigation based reinforcement learning[J]. J of Liaoning Technical University, 2007, 26(6): 915-918.)

[41] Fernandez F, Parker L. Learning in large cooperative multi-robot domains[J]. Int J of Robotics & Automation, 2001, 16(4): 217-226.

[42] Guestrin C, Lagoudakis M, Parr R. Coordinated reinforcement learning[C]. Proc of ICML-02. Sydney, 2002: 227-234.

[43] Jiang D, Wang S, Dong Y. Role-based context-specific multiagent Q-learning[J]. Acta Automatic Sinica, 2007, 33(6): 583-587.

[44] 周浦城, 洪炳镕, 黄庆成. 一种新颖的多 agent 强化学习方法[J]. 电子学报, 2006, 34(8): 1488-1491.
(Zhou P C, Hong B R, Huang Q C. A novel multi-agent reinforcement learning approach[J]. Acta Electronica Sinica, 2006, 34(8): 1488-1491.)

[45] Gu D, Hu H. Fuzzy multi-agent cooperative Q-learning[C]. Proc of ICIA'05. Hong Kong, 2005: 193-197.

[46] 段勇, 崔宝侠, 徐心和. 多智能体强化学习及其在足球机器人角色分配中的应用[J]. 控制理论与应用, 2009, 26(4): 371-376.
(Duan Y, Cui B X, Xu X H. Multi-agent reinforcement learning and its application to role assignment of robot soccer[J]. Control Theory & Applications, 2009, 26(4): 371-376.)

[47] Smith A. Applications of the self-organising map to reinforcement learning[J]. Neural Networks, 2002, 15(8/9): 1107-1124.

[48] Frommberger L. A generalizing spatial representation for robot navigation with reinforcement learning[C]. Proc of AAAI-07. Key West, 2007: 586-591.

[49] Martinson E, Arkin R C. Learning to role-switch in multi-robot systems[C]. Proc of ICRA'03. Taipei, 2003: 2727-2734.

[50] Zheng Liu, Marcelo H Ang Jr. Reinforcement learning of cooperative behaviors for multi-robot tracking of multiple moving targets[C]. Proc of IROS'05. Edmonton, 2005: 1220-1225.

[51] Morales E F. Scaling up reinforcement learning with a relational representation[C]. Proc of the Workshop on Adaptability in Multi-agent Systems. Sydney, 2003: 15-26.

[52] Ghavamzadeh M, Mahadevan S, Makar R. Hierarchical multi-agent reinforcement learning[J]. Autonomous Agents and Multi-agent Systems, 2006, 13(2): 197-229.

[53] Mehta N, Tadepalli P. Multi-agent shared hierarchy reinforcement learning[C]. Proc of the ICML'05 Workshop on Rich Representations for Reinforcement Learning. Bonn, 2005: 45-50.

[54] Whiteson S, Stone P. Concurrent layered learning[C]. Proc of AAMAS-2003. New York: ACM Press, 2003: 193-200.

[55] Mataric M. Learning to behave socially[C]. The 3rd Int Conf on Simulation of Adaptive Behavior. Cambridge: MIT Press, 1994: 453-462.

[56] Price B, Boutilier C. Accelerating reinforcement learning through implicit imitation[J]. J of Artificial Intelligence Research, 2003, 19(1): 569-629.

[57] Jiang C, Sheng Z. Case-based reinforcement learning for dynamic inventory control in a multi-agent supply-chain system[J]. Expert Systems with Applications, 2009, 36(3): 6520-6526.

[58] 李珺, 潘启树, 洪炳熔. 一种基于案例推理的多 agent 强化学习方法研究[J]. 机器人, 2009, 31(4): 320-326.
(Li J, Pan Q S, Hong B R. A CBR-based multiagent reinforcement learning approach[J]. Robot, 2009, 31(4): 320-326.)

[59] Taylor M E, Kuhlmann G, Stone P. Autonomous transfer for Reinforcement learning[C]. Proc of AAMAS'08. Estoril, 2008: 283-290.

[60] 王皓, 高阳, 陈兴国. 强化学习中的迁移: 方法和进展[J]. 电子学报, 2008, 36(12A): 39-43.
(Wang H, Gao Y, Chen X G. Transfer of reinforcement learning: The state of the art[J]. Acta Electronica Sinica, 2008, 36(12A): 39-43.)

[61] Bianchi R A C, Mantaras R L. Should I trust my teammates? An experiment in heuristic multiagent reinforcement learning[C]. Proc of IJCAI'09. Los Angeles, 2009: 11-15.

[62] Buffet O, Dutech A, Charpillet F. Shaping multi-agent systems with gradient reinforcement learning[J]. Autonomous Agents and Multi-agent Systems, 2007, 15(2): 197-220.

[63] Kimura D, Hayakawa Y. Reinforcement learning of recurrent neural network for temporal coding[J]. Neurocomputing, 2008, 71(16/17/18): 3379-3386.

[64] Lagoudakis M G, Parr R. Least-squares policy iteration[J]. J of Machine Learning Research, 2003, 4(12): 1107-1149.

[65] Xu X, Xie T, Hu D W, et al. Kernel least-squares temporal difference learning[J]. Int J of Information Technology, 2005, 11(9): 54-63.

[66] Xu X, Hu D W, Lu X C. Kernel-based least squares policy iteration for reinforcement learning[J]. IEEE Trans on Neural Networks, 2007, 18(4): 973-992.

[67] Busoniu L. Reinforcement learning in continuous state and action spaces[D]. Delft: University Delft, 2008.

[68] Jung T, Polani D. Kernelizing LSPE(λ)[C]. Proc of ADPRL'07. Honolulu, 2007: 338-345.

[69] Ishii S, Fujita H, Mitsutake M. A reinforcement learning scheme for a partially-observable multi-agent game[J]. Machine Learning, 2005, 59(1/2): 31-54.

[70] Wang Y, de Silva C W. A machine-learning approach to multi-robot coordination[J]. Engineering Applications of Artificial Intelligence, 2008, 21(3): 470-484.

[71] Chalkiadakis G, Boutilier C. Coordination in multiagent reinforcement learning: A Bayesian approach[C]. Proc of AAMAS'03. Melbourne, 2003: 709-716.

[72] 郭锐, 吴敏, 彭军, 等. 一种新的多智能体 Q 学习算法[J]. 自动化学报, 2007, 33(4): 367-372.
(Guo R, Wu M, Peng J, et al. A new Q learning algorithm for multi-agent systems[J]. Acta Automatica Sinica, 2007, 33(4): 367-372.)

[73] Tangamchit P, Dolan J, Khosla P. The necessity of average rewards in cooperativemultirobot learning[C]. Proc of ICRA'02. Washington, 2002, 2: 1296-1301.

[74] 任燚, 陈宗海. 基于强化学习算法的多机器人系统的冲突消解策略[J]. 控制与决策, 2006, 21(4): 430-434.
(Ren Y, Chen Z H. Interference solving strategy in multiple robot system based on reinforcement learning algorithm[J]. Control and Decision, 2006, 21(4): 430-434.)

[75] Fan B, Pan Q. Multi-agent coordination based on distributed reinforcement learning and its application to robot soccer[C]. Int Workshop on ETT and GRS'08. Shanghai, 2008: 667-671.

[76] Chang Y H, Ho T, Kaelbling L. All learning is local: Multi-agent learning in global reward games[C]. Proc of NIPS'03. Vancouver, 2003, 16: 807-814.

(下转第1615页)