文章编号:1001 - 9081(2011)03 - 0654 - 03
doi:10.3724/SP.J.1087.2011.00654

基于 π 演算的足球机器人协作 Q 学习方法

柯文德[1,2],朴松昊[2],彭志平[1],蔡则苏[2],苑全德[2]
(1. 广东石油化工学院 计算机科学与技术系,广东 茂名 525000; 2. 哈尔滨工业大学 计算机科学与技术学院,哈尔滨 150001)
(wendeke@163.com)

摘 要:针对多机器人协作学习时出现的学习速度慢、学习效率低等问题,提出了一种基于 π 演算心智模型的足球机器人协作 Q 学习方法,描述了机器人的运动模型,定义了球场现状、目标、意图、行为、协作、请求、扩展知识、能力判断和联合意图等机器人心智状态,构造了联合奖励函数。最后通过实验验证了方法的有效性。

关键词:多机器人;协作;Q 学习;心智状态
中图分类号: TP242.6 **文献标志码**:A

Cooperative Q learning method based on π calculus in robot soccer

KE Wen-de[1,2], PIAO Song-hao[2], PENG Zhi-ping[1], CAI Ze-su[2], YUAN Quan-de[2]
(1. *Department of Computer Science and Technology, Guangdong University of Petrochemical Technology, Maoming Guangdong* 525000, *China*;
2. *School of Computer Science and Technology, Harbin Institute of Technology, Harbin Heilongjiang* 150001, *China*)

Abstract: Concerning the low speed and low efficiency of learning in robot soccer when cooperating between multi-robots, a cooperative Q learning method based on the mental model of π calculus was proposed, in which the mental states were defined as the field state, goal, intention, action, cooperation, request, expanding knowledge, capability judging and connected intention, etc, and the combinational reward function was constructed. The validity of method was verified through experiments.

Key words: multi-robot; cooperation; Q learning; mental state

0 引言

协作问题一直是机器人领域的研究热点之一,其目的是在静态或者动态环境中由若干个同构或异构机器人配合以完成某一共同目标[1]。由于在执行复杂动态协作任务过程中机器人之间经常出现时间冲突、空间冲突与资源冲突,很多学者引入了强化学习方法以解决这些问题,并取得了一些实际效果[2],例如,文献[3]中提出了一种强化学习算法并应用在智能体环境下解决协作问题;文献[4]中提出了一种模糊强化算法并应用在足球机器人双层协作模型上;文献[5]中提出了一种动态环境下的多智能体强化学习协作模型,应用在多机器人协作追捕上等。

尽管目前对强化学习方法在足球机器人比赛环境下的协作应用方面进行了相关研究,但仍然存在一些问题,主要原因在于,当协作环境中的机器人数量增加时,学习空间迅速增大,导致强化学习方法的学习速度下降。考虑到足球机器人比赛环境具有信息不完备的特点,强化学习的联合学习模式具有局限性,特别是多机器人通信和协作时,传统的逻辑方法加入非逻辑性因子以描述通信,并采用问题求解及推理方法以实现协作,不适合多机器人协作的并行性和高实时性要求。由于 π 演算是一种刻画多主体通信系统的进程演算,具有多任务并发执行的特点,其形式化描述手段能够刻画多机器人体系结构的动态性,较好地表示出具有动态结构的进程内以及进程间的交互[6],例如文献[7]基于面向对象的 Petri 网(Object-Oriented Petri nets)和 π 演算,提出一种动态环境下的多 Agent 系统建模方法,较好地满足了多主体的任务交互要求。

基于此,本文在 Q 学习方法的基础上,引入了 π 演算心智模型,体现出数理分析和心智模型推理方法的优点,使足球机器人不但具有强化学习的高度反应能力和环境适应能力,同时具有在动态环境下的推理决策能力。

1 机器人运动描述

多机器人集合 $R = \{R_1, \cdots, R_i, \cdots, R_m\}$ 中,机器人为 3 元组 $R_i = \{\boldsymbol{O}_R, S_R, A_R\}$,其中,$\boldsymbol{O}_R$ 表示机器人的位置与方向,S_R 表示传感器类型,A_R 为任务处理能力。$\boldsymbol{O}_R = [x_i, y_i, \omega_i]^{\mathrm{T}}$,$x_i$、$y_i$ 为第 i 个机器人的位置,ω_i 表示机器人方向,在 t_n 时刻[8]:

$$\begin{cases} x_i^{d_i(t_{n+1})} = x_i^{t_n} + v_i^{t_{n-1}} D_0 \cos(\omega_i + \dot{\omega}_i^{t_{n-1}}) D_0 + \\ \qquad v_i^{t_n} (\Delta t - D_0) \cos(\omega_i + \dot{\omega}_i^{t_{n-1}} D_0 + \dot{\omega}_i^{t_n} (\Delta t - D_0)) \\ y_i^{d_i(t_{n+1})} = y_i^{t_n} + v_i^{t_{n-1}} D_0 \sin(\omega_i + \dot{\omega}_i^{t_{n-1}}) D_0 + \\ \qquad v_i^{t_n} (\Delta t - D_0) \sin(\omega_i + \dot{\omega}_i^{t_{n-1}} D_0 + \\ \qquad \dot{\omega}_i^{t_n} (\Delta t - D_0)) \\ \omega_i = \arctan\left(\dfrac{(y_i(t_n) - y_i(t_{n-1}))}{(x_i(t_n) - x_i(t_{n-1}))} \right) \end{cases} \tag{1}$$

收稿日期:2010 - 09 - 08;**修回日期**:2010 - 11 - 06。 **基金项目**:广东高校优秀青年创新人才培育项目(201180);国家 863 计划项目(2007AA041603;2006AA040202);国家自然科学基金资助项目(60905047;61075076;61075077);国家重点实验室项目(SKLRS200902C);广东省自然科学基金资助项目(8152500002000003)。

作者简介:柯文德(1976 -),男,广东茂名人,副教授,博士研究生,主要研究方向:计算机系统结构、机器人、人工智能; 朴松昊(1972 -),男,黑龙江哈尔滨人,副教授,博士,主要研究方向:计算机软件理论、机器人、人工智能; 彭志平(1969 -),男,福建泉州人,教授,博士,主要研究方向:电子商务、智能主体、机器人; 蔡则苏(1966 -),男,江苏睢宁人,副教授,博士,主要研究方向:计算机软件理论、机器人、人工智能; 苑全德(1981 -),男,山东郓城人,讲师,博士研究生,主要研究方向:并行计算、机器人、智能主体。

其中:$\Delta t = t_{n+1} - t_n$ 为采样时间,D_0 表示传感与计算延迟,$x_i^{d_i(t_n)}$、$y_i^{d_i(t_n)}$ 表示第 i 机器人的坐标位置,ω_i 表示转角角度,$v_i^{d_i(t_n)}$、$\dot{\omega}_i^{d_i(t_n)}$ 分别表示在离散时刻 t_n 的线性速度与角速度。

2　Q 学习算法

Watkins 提出的 Q 学习算法是一种模型无关的动态差分强化学习算法,在无限次遍历状态空间中具有收敛的行动策略,Q 函数构造如下:

$$Q(s,\rho) = \sum_{t=0}^{\infty} \kappa^t E(r_t \mid \rho, s_0 = s) \tag{2}$$

其中:s_0 为初始状态,ρ 为执行策略,r_t 为 t 时刻的奖励,$\kappa \in [0,1]$ 为折扣函数,E 表示折扣奖励期望值,式(2)可改写成:

$$Q(s,\rho) = r(s,a^\rho) + \kappa \sum_{s'} p(s' \mid s,a^\rho) v(s',\rho) \tag{3}$$

其中:a^ρ 为策略 ρ 决定的动作,r 为奖励值,v 为值函数,$p(s' \mid s,a^\rho)$ 表示执行动作 a^ρ 后从状态 s 转移到新状态 s' 的概率。对于离散状态空间 S 中的任何状态 s,其 Bellman 方程为:

$$Q(s,\rho^*) = \max_a \left| r(s,a) + \kappa \sum_{s'} p(s' \mid s,a) v(s,\rho^*) \right| \tag{4}$$

其中 $Q(s,\rho^*)$ 表示在状态 s 下获得的最优值,其近似解为:

$$Q_{t+1}(s,a) = (1-\lambda) Q_t(s,a) + \lambda(\kappa(s,a,s') + \kappa Q_t(s')) \tag{5}$$

其中 λ 表示学习效率。式(5)表明 Q 学习算法必然能够收敛到最优解,通过对动作状态到期望回报映射的动作——评价函数 Q 来解决非马尔可夫问题。

当环境信息完备时,机器人间可通过联合学习以实现任务的协作执行;当环境信息不完备时,机器人执行独立的学习过程,根据所得的奖励来更新维护自身状态—行为对的 Q 值表,每个 Q 值代表执行的优化策略在某个状态—行为对下获得的奖励值,此时进行独立强化学习能够使多机器人协作决策过程收敛[8]。在多机器人系统中,机器人在某个状态下根据 Q 值表执行动作,不断搜索能够取得最大 Q 值的行动策略,直到实现 Nash 平衡。

3　π 演算

π 演算是一种刻画多主体通信系统的进程演算,能较好地表示出具有动态结构的进程内以及进程间的交互[6]。采用 π 演算时,多机器人系统由若干个相互并行的通信与动作进程组成,进程间通过互补链路进行通信。多阶 π 演算的进程如下[6]。

1) 求和。$\sum_{i \in I} P_i = P_1 + P_2 + \cdots + P_n$,表示选择其中的任意一个进程 P_i。

2) 前缀式。$y(\vec{x}) \cdot P$、$\overline{y}(\vec{x}) \cdot P$、$\tau \cdot P$ 分别表示在端口输入/输出名字向量 $\boldsymbol{x}$,或先执行一个不可见动作 τ,然后再执行 P。

3) 组合 $P_1 \mid P_2$。并发地执行进程 P_1、P_2,进程可交换、可结合。

4) 限制 $(\nu y)P$。与进程 P 相似,但名字 y 受到限制,对外界不可见。

5) 匹配 $[\boldsymbol{x} = y]P$。若名字 $\boldsymbol{x}$ 与 y 相同,则执行进程 P,否则结束。

6) 复制 $!P$。提供任意个进程 P 的副本。

7) 结束进程 0。表示进程结束。

8) 通信规则。$(\cdots + \overline{y}(\vec{x}) \cdot P) \mid (\cdots + y(\vec{z}) \cdot Q) \rightarrow P \mid Q\{\boldsymbol{x}/z\}$,其中:向量 $\boldsymbol{x}$ 通过链路 y 在进程间传递,$P \mid Q\{\boldsymbol{x}/z\}$ 表示通信结束后的形式。

4　基于 π 协作演算的 Q 学习

在机器人足球系统中,π 演算通过机器人心智状态演算实现机器人行为的理性和自主性,Q 强化学习通过感知环境、获取奖励并更新 Q 值表执行最优行动策略,通过对环境信息进行推理以实现策略的最优,通过学习不断增强行为效果获取最大收益。

4.1　机器人心智状态

本文在文献[6,9-13]的基础上,提出机器人心智状态的定义。

定义 1　球场现状。$FieldBasedBelief(q) = field(q) \cdot Field_i(q)$,表示第 i 个机器人通过传感器获取球场现状 q,并通过子进程 $Field(q)$ 找到与 q 相关的知识。机器人主观上认为某种状态将要出现,子进程定义为 $FeelBasedBelief(i,\gamma,s) = time(t=\gamma) \cdot [t=\gamma](s(i,\gamma,s) \cdot i(t,s))$,表示在时间 γ,第 i 个机器人处于状态 s。

定义 2　目标。$Goal = goal_i(\vec{q}) \cdot \overline{knowledge}(\vec{q})$,表示接收第 i 个机器人传递来目标 $goal_i$ 时,通过球场 $knowledge$ 查询实现该目标所需要的知识。

定义 3　意图。$Intention = intend(\gamma,\xi) \cdot \overline{time}(\gamma) \mid \overline{s}(i,\gamma,\vec{\xi}) \mid \overline{goal}(\vec{\xi})$,表示 γ 时刻机器人通过传感器感知球场状态蕴含的意图 ξ,判断是否能够实现目标并提出请求。

定义 4　行为。$PerceptionPlan(\varepsilon) = perceive(\gamma,\varepsilon) \cdot (\overline{intend}(\gamma,\xi) \mid ActionPlan(\xi))$,表示机器人在时间 γ 感知到外界刺激 ε 后,查询所蕴含的意图 ξ 并输出目标 ξ,同时规划行动以实现目标。

定义 5　协作。$Offer(w,u) = !S(w,r) \cdot (\overline{w}(u) \mid \overline{field}(u))$,表示当机器人接收到协作请求 r 与目标位置 w 后,向请求方输出协作 u,并查询实现该协作所需的知识。

定义 6　请求。$Request(w,r) = r(w,r) \cdot w(c)$,表示机器人输出协作请求 r 并确定目标位置 w 后,提供协作 c。

定义 7　扩展知识。$ExpandKnowledge(e,k) = expand(i) \cdot Expand(e,k)$,表示接收其他机器人提供的知识 k 后,与原有的知识 e 合并扩展。

定义 8　能力判断。$Capability = [x = \varepsilon]((\overline{field}(x) \mid FieldBasedBelief(q)) \cdot Renew(q))$,表示接收到球场环境刺激 ε 后,若自身能力能够提供输出,则更新能力,且不再提出请求。

定义 9　联合意图。

$ConnectedIntention(\gamma,I_j) = cintention(\gamma,I_j) \cdot ((believe(i -$

$1,\gamma,intention_{i-1}) \mid believe(i,\gamma,intention_i) \mid believe(i+1,\gamma,intention_{i+1}))\cdot cintention(\gamma,intention_j)$

表示第 j 个机器人相信其他 $(i-1,\cdots,i+1)$ 机器人将会在规定时间 γ 内完成各自的子任务，为联合意图作出的承诺完成共同的目标。

4.2 Q 学习的奖励函数

机器人离散动作模型中的初始 $Q(k,a)$ 为任意值，其中 $k \in S$、$a \in A$，根据 Q 值和环境状态随机生成动作行为序列，采用 Borlzman 分布确定随机动作的随机度[5]：

$$p(a_i) = \exp[Q(\xi,a_i)/T] \sum_{a_\xi \in A} \exp[Q(\xi,a_\xi)/T] \tag{6}$$

其中随机性随温度值 T 的增加而增大。

在 π 演算心智模型下，定义协作成功的奖励值为 μ，协作不成功的奖励值为 $-\mu$，奖励函数 $s_r \in [-1,1]$，阈值取 0.095μ。在实现最终目标后，可以通过反向反馈计算出奖励函数。当机器人协作成功时，奖励函数 $R=\mu$，否则 $R=-\mu$。

定义机器人向足球移动的奖励函数如下：

$$r^1 = \begin{cases} \tau \mid x^h - x_t \mid, & \mid x^h - x_t \mid \leqslant \sigma \\ -\tau \mid x^h - x_t \mid, & \text{其他情况} \end{cases} \tag{7}$$

其中：τ 为奖励系数；x_t 表示在 t 时刻足球的坐标；x^h 表示我方机器人的坐标位置；σ 为阈值距离，当有效机器人与足球的距离达到阈值以内时，机器人得到奖励。

定义机器人将球传给队员的奖励函数如式(8) 所示：

$$r^2 = \begin{cases} \tau \mid x^h - x_t \mid, & \text{传球成功} \\ -\dfrac{\tau}{\mid x^o - x_t \mid}, & \text{传球失败} \end{cases} \tag{8}$$

其中 x^h、x^o 分别代表我方接球队员与对方截球队员。

定义机器人将球射入球门的奖励函数如下：

$$r^3 = \begin{cases} \eta, & \text{将球射入对方球门} \\ -\eta, & \text{将球射入我方球门} \\ 0, & \text{其他情况} \end{cases} \tag{9}$$

其中 η、$-\eta$ 分别表示正向奖励和负奖励。

从而，按照加权的方法得到如下的联合奖励函数：

$$r = \alpha_1 r^1 + \alpha_2 r^2 + \alpha_3 r^3 \tag{10}$$

其中：α_1、α_2、α_3 为对应的加权系数，且 $\alpha_1,\alpha_2,\alpha_3 \geqslant 0$，$\alpha_1+\alpha_2+\alpha_3=1$。

5 实验验证

分别在 FIRA Simurosot 5VS5 仿真平台和 MOS2007 仿人机器人平台上对本方法进行验证。图 1 所示为一次成功的协作，左方队员(R1、R2、R3)突破右方队员(O1、O2、O3、O4、O5)的防守，将球踢进对方球门的场景，图 2 所示为 MOS2007 仿人机器人的一次成功协作射门的场景，图 3 比较了传统的 Q 学习方法和基于 π 协作演算的 Q 学习的成功次数，可以看到，基于 π 协作演算的 Q 学习的成功次数明显高于传统的 Q 学习方法。

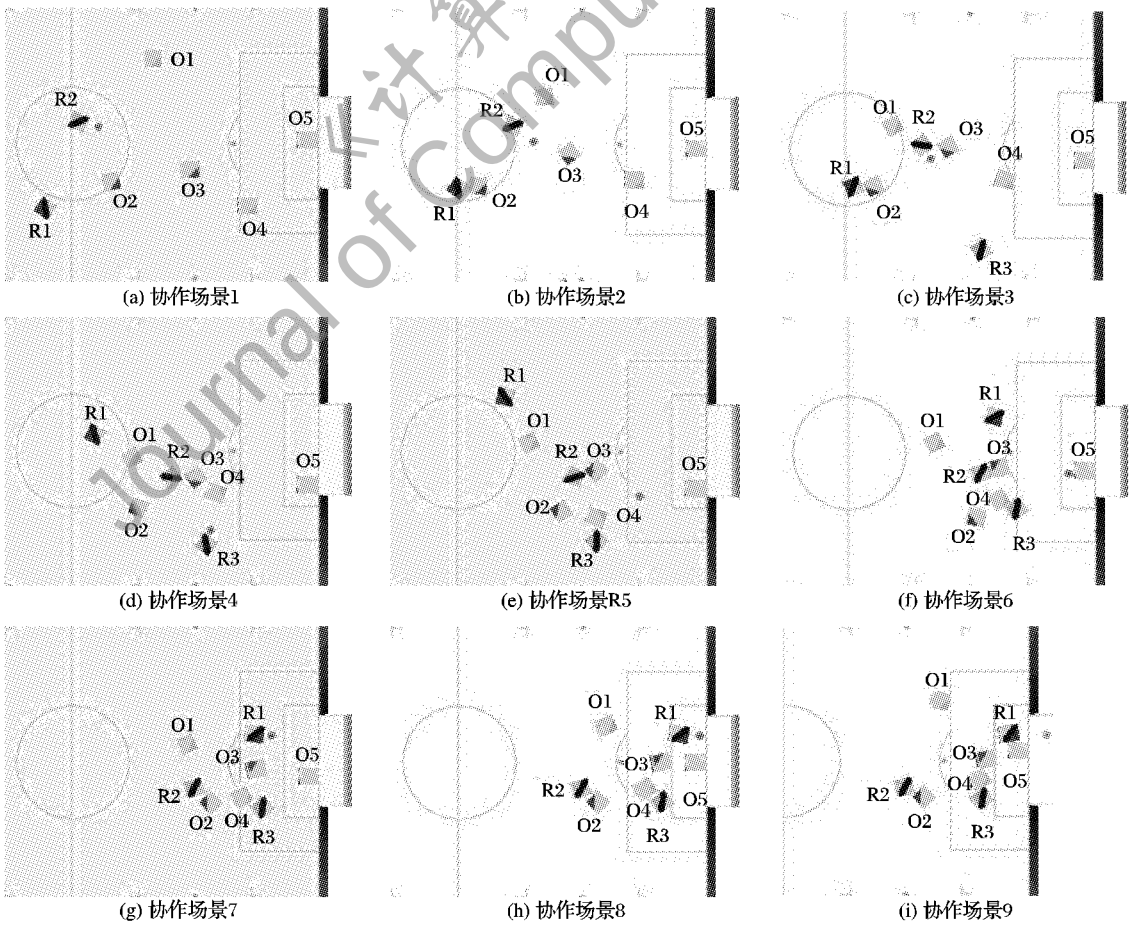


图 1 仿真 5VS5 机器人平台上一次成功的协作踢球

(下转第 669 页)

反观图 5,从图中可以看出综合批号不但出现了图 4 中的综合批号分配错误、对调、出现新批的现象,而且还出现了新的比较严重的错误。综合批号 4 从 5 s 时就消失了,综合目标 4 对应的真实目标的综合批号由 4 改为了 3,综合批号 3 对应的真实目标在 5 s 时开始综合批号已改为 6,这样将错误地认为综合批号 4 代表的目标消失,并出现新目标即综合批号 6 代表的目标,其实综合批号 6 代表的目标是真实目标 3,综合批号 3 在不同时刻代表的不是同一目标,从而无法对目标进行有效地跟踪及识别。

通过这些仿真结果可以得出,本文提出的基于渴望度的关联航迹自动编批算法很好地反映了关联的结果,并为局部航迹分配综合批号,在出现关联错误的情况下,目标的综合批号虽然不可避免地发生了跳动,但总体而言,多数目标的综合批号未变,综合批号的跳动不大,同一综合批号在不同时间尽可能地代表同一目标,基本实现了局部航迹与系统航迹的对应,因此该算法对关联错误有一定的容忍能力。而未利用历史信息的编批算法仿真的结果则不理想,关联错误严重地影响了该算法的性能,导致无法实现局部航迹与系统航迹的对应,不利于后续的航迹融合、目标跟踪和识别。

4　结语

本文针对存在关联错误时编批算法遇到的困难,提出了一种基于渴望度的关联航迹自动编批算法。该算法利用聚类结果、渴望度及局部航迹的综合批号历史信息为每一聚类分配综合批号,并实现关联的局部航迹与系统航迹的对应,以便后续的航迹融合、目标跟踪和识别。仿真结果表明依据渴望度分配综合批号可以使综合批号的跳动最少,将关联错误引起的影响降到最低,使综合批号在不同时间尽可能地代表同一目标。该算法适用于环境复杂和恶劣的实际工程系统,具有较好的稳健性。

参考文献:

[1] 韩崇昭, 朱洪艳, 段战胜, 等. 多源信息融合[M]. 北京: 清华大学出版社, 2006.

[2] 徐毅, 金德琨. 数据融合体系结构的设计[J]. 航空电子技术, 2001, 32(4): 25 - 31.

[3] BAR-SHALOM Y, FORTMANN T E. Tracking and data association [M]. New York: Academic Press, 1988.

[4] 石玥, 王钺, 王树刚, 等. 基于目标参照拓扑的模糊航迹关联方法[J]. 国防科技大学学报, 2006, 28(4): 105 - 109.

[5] STONE L D, WILLIAMS M L, TRAN T M. Track to track association and bias removal [C]// Proceedings of the SPIE Conference on Signal and Data Processing of Small Targets. Orlando, FL: [s. n.], 2002: 315 - 329.

[6] 黄友澎, 周永丰, 张海波, 等. 一种多雷达航迹加权融合的权值动态分配算法[J]. 计算机应用, 2008, 28(9): 2452 - 2454.

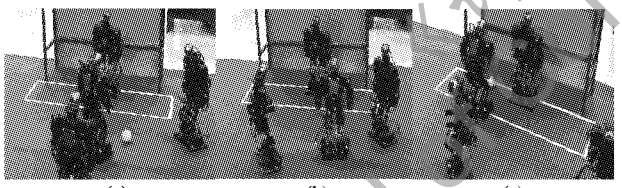
[7] 朱洪艳, 韩崇昭, 韩红, 等. 航迹起始算法研究[J]. 航空学报, 2004, 25(3): 284 - 288.

[8] 徐盐生. 滑窗式航迹自动起始[J]. 舰船电子对抗, 1998(2): 23 - 25.

[9] 朱洪艳, 韩崇昭, 韩红. 基于期望极大化算法的航迹起始方法研究[J]. 计算机工程与应用, 2003, 39(14): 66 - 69.

[10] 毛明秀. 雷达多目标自动编批研究[D]. 南京: 南京理工大学, 2007.

(上接第 656 页)



(a)　(b)　(c)

图 2　MOS2007 仿人机器人平台上一次成功的协作踢球

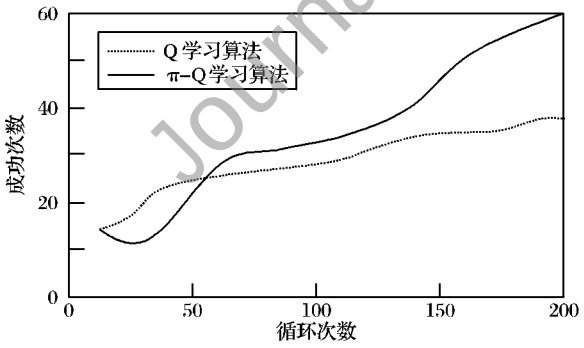


图 3　Q 学习方法与基于 π 演算的 Q 学习方法比较

6　结语

本文提出了在信息不完备的动态环境下,将机器人的 Q 学习与 π 演算模型结合起来,体现出数理分析和心智模型推理方法的优点,通过在足球机器人比赛中的应用,使协作模型不仅具备强化学习的高度反应性和自适应性,并拥有 π 演算的推理能力,实验的比赛结果表明了该方法的有效性。

参考文献:

[1] 石志国, 王志良, 刘冀伟, 等. 基于周期时间限制的多机器人自主委托协作模型[J]. 机器人, 2010, 32(1): 109 - 118.

[2] 高阳, 陈世福, 陆鑫. 强化学习研究综述[J]. 自动化学报, 2004, 30(1): 86 - 100.

[3] 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.

[4] 曹卫华, 徐凌云, 吴敏. 模糊 Q 学习的足球机器人双层协作模型[J]. 智能系统学报, 2008, 3(3): 234 - 238.

[5] 朴松昊, 孙立宁, 钟秋波, 等. 动态环境下的多智能体机器人协作模型[J]. 华中科技大学学报: 自然科学版, 2008, 36(z1): 39 - 41.

[6] 史忠植. 智能主体及其应用[M]. 北京: 科学出版社, 2000.

[7] 于振华, 蔡远利, 徐海平. 基于 π 网的多 Agent 系统建模与分析[J]. 系统工程理论与实践, 2007, 27(7): 77 - 84.

[8] HARMATI I, SKRZYPCZYK K. Robot team coordination for target tracking using fuzzy logic controller in game theoretic framework [J]. Robotics and Autonomous Systems, 2009, 57(1): 75 - 86.

[9] 杨鲲, 翟永顺, 刘大有. Agent: 特性与分类[J]. 计算机科学, 1999, 26(9): 30 - 34.

[10] 陈为雄, 李振龙. 基于 BDI 模型的多机器人智能体系统设计[J]. 机器人, 2004, 26(4): 310 - 313.

[11] 康辉, 曾莹莹, 刘志勇. 基于 PI-演算的移动通信服务研究与建模[J]. 通信学报, 2009, 30(4): 11 - 16.

[12] 廖军, 谭浩, 刘锦德. 基于 PI-演算的 Web 服务组合的描述和验证[J]. 计算机学报, 2005, 28(4): 635 - 643.

[13] 李长云, 李赣生, 何频捷. 一种形式化的动态体系结构描述语言[J]. 软件学报, 2006, 17(6): 1349 - 1359.