文章编号: 1002-0446(2000)06-0482-08

机器人足球赛中基于增强学习的任务分工*

顾冬雷　陈卫东　席裕庚

(上海交通大学自动化研究所　上海　200030)

摘　要: 本文研究了机器人足球赛中利用增强学习进行角色分工的问题, 通过仿真试验和理论分析, 指出文[1]中采取无限作用范围衰减奖励优化模型(infinite-horizon discounted model)的 Q 学习算法对该任务不合适, 并用平均奖励模型(average-reward model)对算法进行了改进, 实验表明改进后学习的收敛速度以及系统的性能都提高了近一倍.

关键词: Q 算法; 无限作用范围衰减奖励优化模型; 平均奖励优化模型

中图分类号:　TP24　　文献标识码:　B

1　引言

利用各种学习方法提高移动机器人的适应能力是机器人领域当前的研究热点. 作为一种无监督的学习方法, 增强式学习由于它的普遍适用性受到了广泛的关注. 其中 Q 算法[2]由于其实现的简易性, 在研究中得到了最广泛的使用.

文[1]利用 Q 算法[2]研究了机器人足球赛中的角色分工. 文中假设每个机器人的行动能力、感知能力完全一样, 基本行为也完全一样, 一方由预编程决定前锋后卫的角色分工, 另一方通过学习来确定机器人的角色分工. 最后学习一方的性能超过了预编程的一方.

本文在我们自行开发的机器人足球赛仿真系统中仿真了文[1]中的实验. 在实验过程中发现, 利用 Q 学习算法, 学习结果和诸多因素有关, 很难达到文[1]中所记录的性能. 通过分析, 本文认为增强学习在该实验中的任务是学到优化的静态角色分工, Q 方法采用的无限作用范围衰减奖励策略优化模型(infinite-horizon discounted model)[3]常用于学习优化的动态行为序列, 不符合本实验的实际情况, 而平均奖励策略模型(average-reward model)[3]比较有效. 仿真实验表明, 平均奖励模型的使用, 加快了学习过程的收敛性, 同时使系统性能提高了近一倍.

2　机器人足球赛中基于增强学习的任务分工

在文[1]中, 机器人足球赛的每支队伍由 4 个机器人组成, 球场的边界是墙壁, 足球无法出界, 只能弹回. 将整个底线当作球门, 只要球触及了对方的底线, 即认为进球得分. 比赛是连续的, 即任何一方得分后, 足球立即被放置到球场的中心, 进行下一轮的较量. 为了简化学习过程, 比赛没有时间限制, 双方得分和达到一定数值时比赛结束. 机器人足球队性能的确定为己方的净胜球数 $P = S_{us} - S_{them}$, 其中 S_{us} 与 S_{them} 是各自队伍的得分情况.

为了方便比较分析, 每个机器人的传感器能力与行动能力都相同. 每个机器人固有的基本行为也是相同的. 文中定义了三种行为集合:

---

* 基金项目: 本项目获 863 项目(863-512-9805-18) 及国家自然科学基金(69889501)资助.
收稿日期: 2000-01-27

(1) 向足球移动(m tb): 机器人直接向球运动, 碰撞促使球离开机器人.

(2) 包抄至球后面(gbb): 机器人一边躲开和球的碰撞, 一边包抄到球后面与己方底线的某个位置上.

(3) 运动到后场(m tbf): 机器人在己方后场运动, 同时适时地采取进攻行为.

系统的整体行为通过根据感知到的情况有选择地激活三种行为之一来体现. 每个机器人感知到的情况被简化成感知自身是否在足球与己方底线之间(bb), 据此来选择三种基本行为之一. 于是, 有以下 9 种机器人策略:

表 1　机器人可选的策略

	M tb	gbb	m tbf	M tb	gbb	m tbf	M tb	gbb	m tbf
Not bb	0	0	1	0	0	1	0	0	1
bb	0	0	1	0	1	0	1	0	0
Not bb	0	1	0	0	1	0	0	1	0
bb	0	0	1	0	1	0	1	0	0
Not bb	1	0	0	1	0	0	1	0	0
bb	0	0	1	0	1	0	1	0	0

文[1]让一队机器人由程序预先决定其任务分工, 另一队通过学习决定如何分工. 预定分工的足球队的策略为:

表 2　进攻策略

感知情形	采取动作		
	m tb	gbb	m tbf
Not bb	0	1	0
bb	1	0	0

表 3　防守策略

感知情形	采取动作		
	m tb	gbb	m tbf
Not bb	0	1	0
bb	0	0	0

其中 3 个机器人采用进攻策略, 扮演前锋角色, 一个机器人采用防守策略, 扮演守门员角色.

另外一队则采用 Q 学习算法来让足球队学到角色分工. 算法完全按照文献[2]中的算法, 其递归步骤如下:

(1) 观察当前状态 x_n;

(2) 选择并且执行行为 a_n;

(3) 观察后续的状态 y_n;

(4) 得到立即的奖励 r_n;

(5) 更新 Q_{n-1} 的值.

$$Q_n(x,a)=\begin{cases}(1-\alpha_n)Q_{n-1}(x,a)+\alpha_n[r_n+\gamma V_{n-1}(y_n)] & \text{if } x=x_n \text{ and } a=a_n\\ Q_{n-1}(x,a) & \text{otherwise}\end{cases} \tag{1}$$

其中

$$V_{n-1}(y)\equiv \max_b\{Q_{n-1}(y,b)\} \tag{2}$$

(6) 重复以上过程.

采用的增强函数为

$$R(t)=\begin{cases}1, & \text{our team scored}\\ -1, & \text{other team scored}\\ 0, & \text{otherwise}\end{cases} \tag{3}$$

学习在每一个迭代步骤上进行. 以双方得分和满 10 分为一回合, 连续 100 回合作为一次

实验, 进行了 10 次实验, 文中的实验结果如下:

性能: 通过学习决定分工的队伍胜过固定策略的队伍, 每回合平均比分为 6: 4, 其中包括了没有学习到优化策略的初始情形.

策略收敛: 采用观察机器人策略变化的频率来考查过程. 每个机器人策略变化在 100 回合游戏后平均每回合变化 0. 25 次.

3 仿真研究及对任务分工中增强算法的改进

本文开发了相应的仿真环境, 进行了仿真研究.

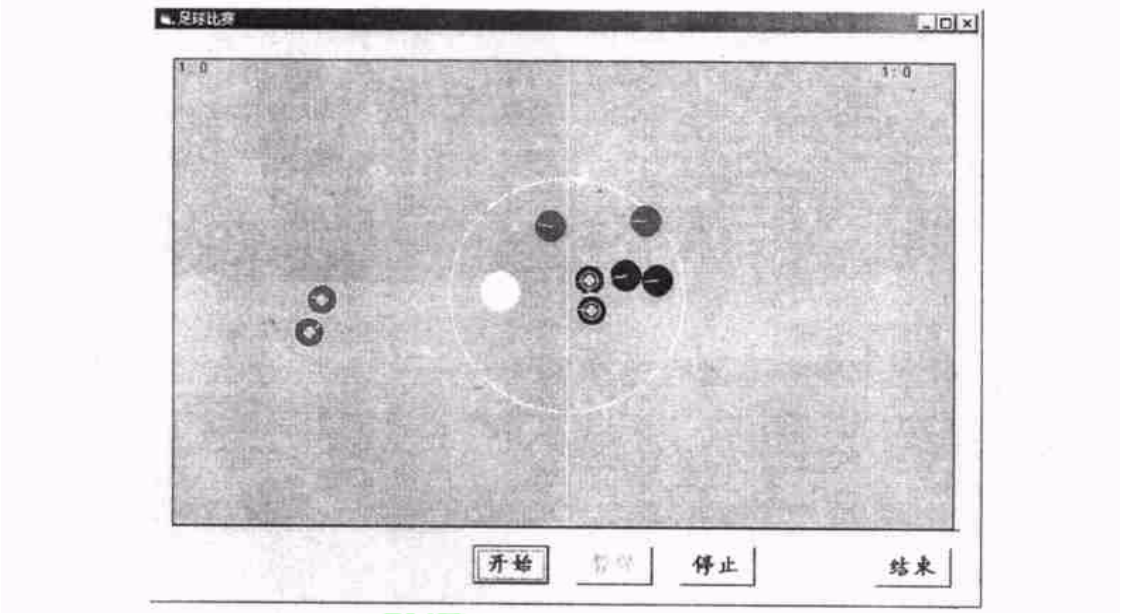

图 1 机器人足球赛仿真系统界面

为了研究 Q 算法的有效性, 首先考察了完全随机挑选策略的球队和预先决定分工球队的比赛. 预先决定分工球队的策略同文[1], 见表 2、表 3. 以总比分和 100 分为一次实验, 进行了 10 次实验, 平均比分为 92. 3: 7. 7, 明显地, 采用随机策略的队伍无法和预先分工的队伍抗衡.

然后我们检验了 Q 算法对性能的影响. 因为 Q 算法选取行为时采取的是贪心法, 每次总是选取最大化 $Q(s, a)$ 值的行为, 容易陷入局部最优值, 因此在行为选择时引入一定概率的随机变化, 让机器人在用贪心法选择行为的同时能以一定概率探索不同的策略. 随着训练的进行, 逐渐以几何速度减少随机性, 让算法逐渐收敛到优化值. 因为仿真采用的是伪随机数, 实际上整个过程是确定性的, 因此, 每次实验的随机函数都采用了不同的种子值作随机函数的参数, 使每次实验的过程都不相同.

以双方进球数和为 10 分作为一个比赛回合, 以连续 50 个回合作为一次实验, 做了 10 次实验. 实验结果表明, 系统性能对 Q 算法的初始值有比较明显的依赖性. 当将初始值全部设置为 0 时, 一般而言, 学习方的进球总数至多是接近预先分工的一方, 当然和采用完全随机的情况相比, 学习的效果是相当明显的. 但是想要达到文章[1]中所述的性能比较困难.

以一次实验为例. 仿真步长为 0. 5s. Q 算法的 α 值取 0. 01, γ 取值 0. 3, 开始时引入概率为

0. 3 的随机行为选择, 此概率以 0. 995 的几何速度衰减. 比赛结果的学习方和固定分工方的总比分为 215: 285, 学习方没有能够胜过固定分工方.

图 2 是学习方各个机器人控制器中学习算法 Q 值变化趋势图. 此处并未将所有 Q 值记录下来, 只是将在得分与失分时的 Q 值加以记录, 以便和后面的改进方法进行比较. 可以看见, Q 值收敛性比较差, 直到仿真过程的最后阶段, 还有许多不稳定的因素. 图 3 是随着比赛进行, 每回合机器人改变自己策略的次数. 如果改变策略次数减少到 0, 那么策略就完全收敛了. 从图中可以见到, 50 回合比赛结束时, 机器人的策略还没有形成比较平坦的曲线, 说明收敛的速度比较慢, 尚未进入收敛区域. 图 4 是每一回合比赛学习方的净得分数的变化. 可以看到, 随着比赛的进行, 学习方的净得分值逐渐增大, 性能逐渐赶上固定分工的一方.

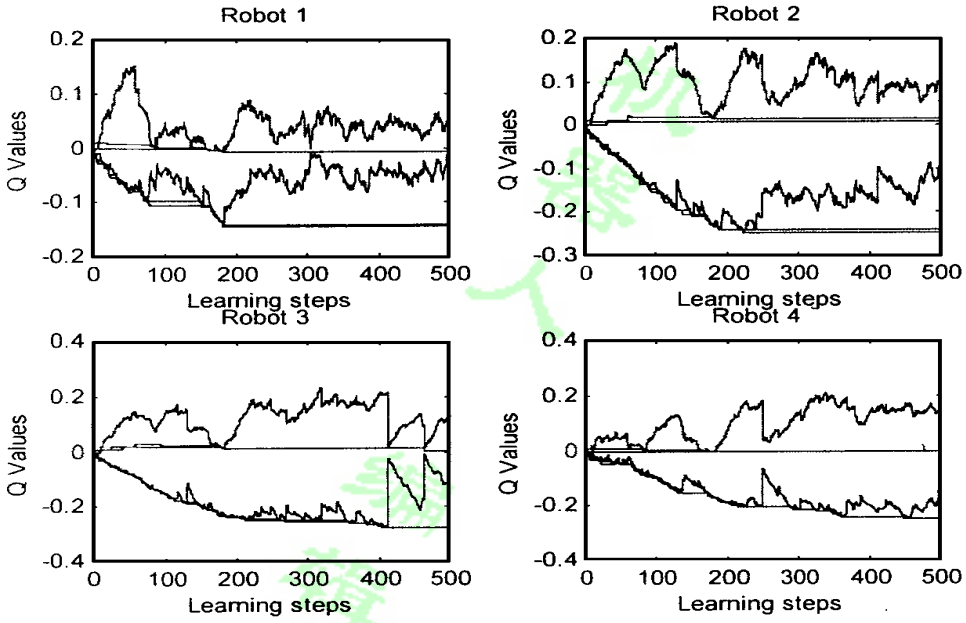

图 2　Q 值变化趋势图

实验表明, 文[1]采用 Q 算法决定角色分工的实际效果并不理想.

一般地, 考虑采用增强学习算法使行为优化之前, 必须首先决定采用什么样的优化模型, 即必须决定智能体如何把将来的效用考虑进它当前要执行的行为中去[3]. 当前该领域中有三种典型的模型:

有限作用范围模型(finite-horizon): 在一个给定的时刻, 智能体必须最优化以下 h 步的期望奖励值. 即

$$E[\sum_{t=0}^{h} r_t] \tag{4}$$

而不管后面将发生什么. 此处 r_t 代表将来的第 t 步接收到的奖励. 这个模型一般用于智能体的策略不稳定的情况, 即随时间变化的情况. 在它采取第一步行为的时候, 假定进行一个 h 步优化策略的行为. 第二步则是进行一个 $h-1$ 步的优化策略行为, 以此类推直到智能体采取一步优化策略然后结束.

无限作用范围衰减模型(infinite-horizon discounted model) 考虑智能体长远的得到奖励

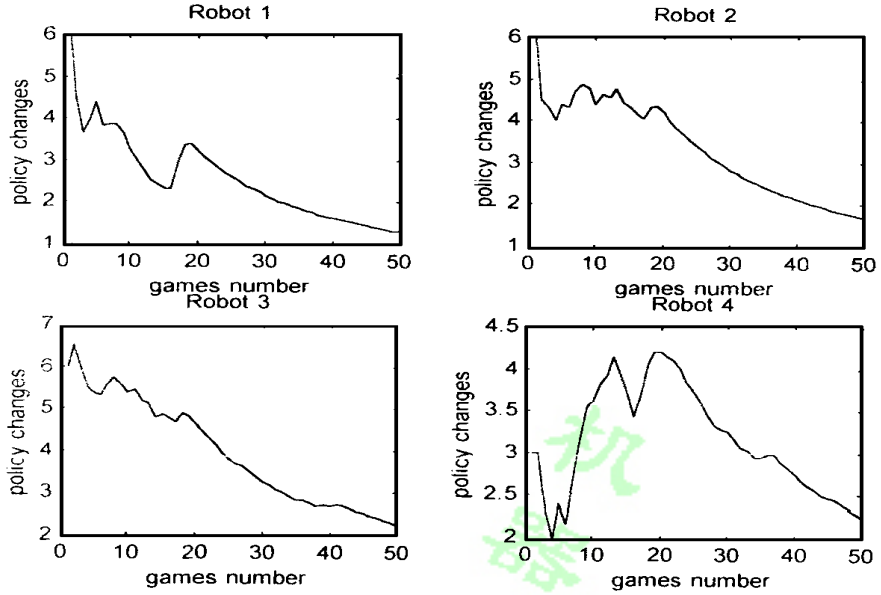

图 3 每比赛回合策略改变次数

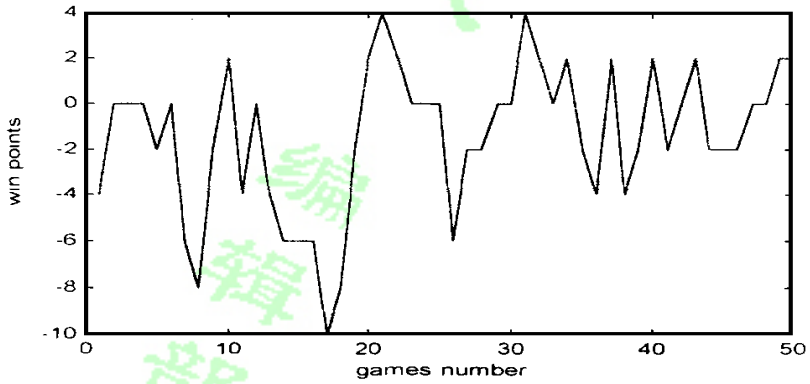

图 4 每局胜负情况

的情况, 但是将来得到的奖励总是由一个衰减因子 $\gamma(0 \leqslant \gamma < 1)$ 进行几何衰减.

$$E[\sum_{t=0}^{\infty} \gamma^t r_t] \tag{5}$$

γ 的意义可以有多种方法加以解释. 一般认为, 当前行为的选择对将来的每一步都起作用, 但是作用大小是逐渐衰减的, γ 是衰减率. 通常用于动态连续决策的场合, 用以获取一个优化的行为序列. γ 的引入使它在数学上确保无限次累加和有界, 这使它在数学上更加容易处理, 也是它获得广泛关注的主要原因.

另外一个优化策略是平均奖励模型(average-reward model), 智能体考虑采取使它的长期平均奖励最优化的行动

$$\lim_{h \to \infty} E(\frac{1}{h} \sum_{t=0}^{h} r_t) \tag{6}$$

文[2]中 Q 算法采用了无限作用范围奖励衰减模型, 对于这个机器人足球赛的研究任务, 它是不合适的. 虽然机器人足球比赛本身是一个动态过程, 但是, 因为所研究的是机器人采用的分工策略和球队性能之间的关系, 学习方球队的得分状况完全依赖于它们采用的任务分工, 学习的目的是学到一种静态的角色分配关系, 而并非是一个动态的行为序列. 而采用无限范围奖励衰减模型针对整个过程进行学习, 主要是用来学习得到一个优化的动态连续决策序列, 显然不符合这个要求.

考察另外两种模型, 第一种模型明显是不符合要求的, 不予考虑. 但平均奖励模型有明显的意义, 因为得分对应着奖励 1, 失分对应着惩罚 - 1, 那么平均奖励模型和任务的性能指标—— 平均净得分数—— 完全一致. 增强函数和优化模型之间的关系是完全一致的, 有明确的意义. 进而, 在以相同的任务分工对待预先确定分工的一方, 并不是每次对局都能得到相同的结果的, 因为得分与否, 队形仅仅是因素之一, 其他一些动态交互的因素由于太复杂, 文[1]中没有加以考虑, 因此, 某种角色分工的得分情况应该是一个概率分布. 这样, 平均奖励模型的平滑特性也无形之中对此加以了考虑.

基于这些考虑, 本文采用了平均奖励优化模型代替了 Q 算法采用的无限作用范围奖励衰减模型, 算法流程如下:

(1) 观察当前状态 x_n;

(2) 选择并且执行行为 a_n;

(3) 观察后续的状态 y_n;

(4) 得到立即的奖励 r_n;

(5) 更新 Q_{n-1} 的值.

$$Q_n(x,a) = \frac{P_{n-1}(x,a) \times Q_{n-1}(x,a) + r_n + V_{n-1}(y_n)}{P_{n-1}(x,a) + 1} \tag{7}$$

$$P_n(x,a) = P_{n-1}(x,a) + 1 \tag{8}$$

其中

$$V_{n-1}(y) \equiv \max_b\{Q_{n-1}(y,b)\} \tag{9}$$

$P_n(x,a)$ 是进行第 n 步训练时在状态 x 执行动作 a 的次数.

(6) 重复以上过程.

根据这个算法, 进行了新的实验, 所有其他相关参数完全和前述的实验相同, 但是训练只在得分或者失分的时刻进行, 而在整个进攻防守动态过程中不进行学习, 这是因为所要学习的是一种静态的优化的角色分工, 而非优化的进攻、防守动态过程. 于是, 增强函数改变成

$$R(t) = \begin{cases} 1, & \text{our team scored} \\ -1, & \text{other team scored} \end{cases} \tag{10}$$

以某一次实验为例, 随机函数种子值挑选成 54321. 由图 5 所示, 200 步时, Q 值基本收敛到稳定值上. 由图 6 可以看到, 策略的收敛速度相当快, 比赛进行到 50 回合时, 变化次数就已经减少到每回合变化 0.1 次左右, 远快于文[1]中所述的 100 回合后收敛到每回合变化次数 0.25. 由图 7 表明, 3 回合比赛之后, 学习方的性能就已经超过了预先决定分工角色的一方. 最后学习方和固定分工方的总比分为 379:121, 平均比分约为 6:2, 比文[1]报道的 6:4 的平均比分性能提高了一倍.

改变随机函数种子值, 重复进行多次实验, 显示出了良好的可重复性. 而用这种策略优化

模型进行学习, 无需调节学习参数, 比起采用无限作用范围奖励衰减模型, 学习过程得到了简化.

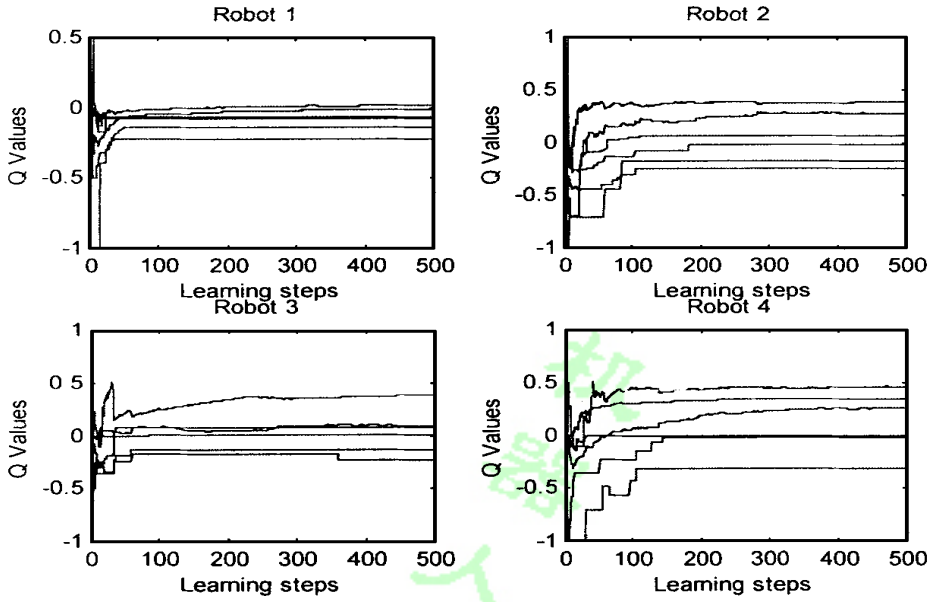

图 5 平均奖励策略模型 Q 值收敛情况

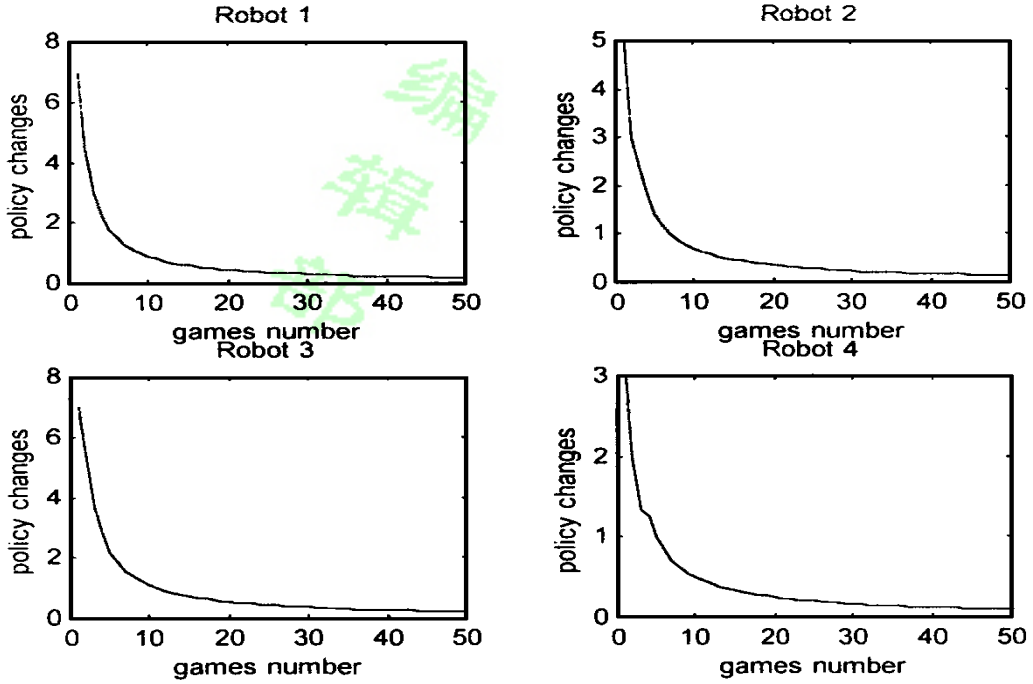

图 6 平均奖励策略模型每局策略改变

4 结论

学习算法应用于移动机器人时, 注意力往往集中在如何写出比较好的增强函数, 如何确保

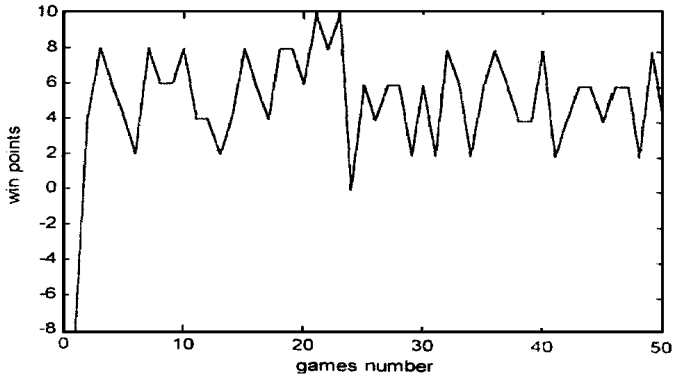

图 7　平均奖励模型每局胜负情况

算法的收敛性等方面. 考虑究竟要优化什么指标则较少. 由于无限作用范围奖励衰减模型的形式在数学上最容易加以处理从而得到比较好的结论, 因此应用最为广泛. 但是对于机器人足球赛中静态的角色分工问题, 采用平均奖励模型更加符合实际情况. 本文通过理论分析改进了算法, 其有效性在仿真实验中得到了验证.

参　考　文　献

1　Tucker Balch. Learning Roles: Behavioral Diversity in Robot Teams. In AAAI-97 Workshop on Multiagent Learning, Providence, R. I, 1997

2　Christopher J C H Watkins. Technical Note: Q-Learning. Machine Learning, 1992, 8: 279- 292

3　Leslie Pack Kaelbling, Michael L Littman. Reinforcement Learning: A Survey. Journal of Artificial Intelligence Research 1996, 4: 237- 285

ROLE DIVERSITY IN ROBOT SOCCER BASED ON REINFORCEMENT LEARNING

GU Dong-lei　CHEN Wei-dong　XI Yu-geng

(*Institute of Automation, Shanghai Jiaotong University*　200030)

Abstract: In this paper, the role diversity based on reinforcement learning in robot soccer is studied. Through simulation and analysis, it is shown that the Q algorithm infinite-horizon discounted model in [1] is not suitable to this task. Instead of that, average-reward model is used for improving the algorithm. Simulation experiments show that the convergence rate in learning and the system performance are twice increased after improvement.

Keywords: Q algorithm, infinite-horizon discounted model, average-reward model

作者简介:

顾冬雷 (1971-), 男, 博士生. 研究领域: 智能机器人, 多机器人协调控制.

陈卫东 (1967-), 男, 博士, 副教授. 研究领域: 机器人, 多智能体协调控制.

席裕庚 (1944-), 男, 教授, 博士生导师. 研究领域: 预测控制, 多智能体协调控制等.