文章编号: 1002-0446(2003)07-0761-06

仿真机器人足球中的强化学习

宋志伟, 陈小平
(中国科学技术大学计算机科学技术系, 安徽 合肥 230027)

摘 要: 本文总结当前仿真机器人足球中强化学习的研究进展,系统阐述在仿真机器人足球不同决策层次中使用强化学习的不同方法,针对仿真机器人足球的特点讨论当前使用的几种对环境状态空间进行泛化的方法,并展望今后强化学习在仿真机器人足球中的主要应用方向.

关键词: 仿真机器人足球; 强化学习; 多主体系统
中图分类号: TP24 **文献标识码**: B

REINFORCEMENT LEARNING IN ROBOCUP SIMULATION

SONG Zhi-wei, CHEN Xiao-ping
(*Computer Science Department, University of Science & Technology of China, Hefei 230027, China*)

Abstract: This paper summarizes the recent research progress of reinforcement learning used in RoboCup simulation, and at the same time states the different methods used in different decision-making layer. This paper also discusses the commonly used generalization methods for the large state space and then gives a prospect for the application of reinforcement learning in RoboCup simulation.

Keywords: RoboCup simulation; reinforcement learning; multi-agent systems

1 引言(Introduction)

仿真机器人足球(RoboCup Simulation)是多主体系统(Multi-agent Systems)研究的一个标准问题,同时包含了动态、实时、不确定环境中的合作与对抗,已被国际人工智能界公认为人工智能新的挑战问题[1]. 在这种系统的设计中,完全依靠手工编码是十分困难的,设计者很难把各种情况都考虑清楚. 最好能使足球机器人象人一样,通过不断的学习和训练,学会踢球. 为此,很多研究者尝试了利用机器学习,特别是强化学习(Reinforcement Learning)技术,通过主体与环境的交互和试错,不断增强其自身能力.

由于仿真机器人足球的复杂性,难以将强化学习的方法直接加以应用. 人们一般先要把主体的决策分成几个不同的层次,在每个层次中又分成一些不同的任务,对应每个任务再使用相应的强化学习或其他学习方法来完成学习任务. Peter Stone 率先提出了分层学习(Layered Learning)的想法,并且给出了四个原则[2]. 作者结合人类足球的特点,进一步给出了仿真机器人足球中的一种分层决策(Layered Decision-making)模型[3],把仿真机器人足球中主体的决策分为三个层次,分别是: 个人技术决策(Individual Skills)、局部战术决策(Tactics)和全局战术决策(Strategy). 个人技术决策是最低层的决策,它的任务是一些个人必备的技术,如传球、带球、射门等,它们的决策结果是由基本动作组成的动作序列; 局部战术决策是较高层次的决策,它的任务主要是几个队友之间的一些战术配合,如 2 过 1 等,决策结果是个人技术层的任务(即个人技术)组成的序列; 全局战术决策是最高层的决策,涉及阵型、战术风格等,决策结果是战术配合组成的序列.

在全局战术层,目前还没有很好的学习方法,通常仍依靠手工编码. 在个人技术层和局部战术层,强化学习使用得比较多,并且取得了值得重视的进展. 本文第二节简单介绍强化学习的基本内容; 第三节和第四节分别介绍个人技术层和战术层的强化学习; 最后第五节是结论和展望.

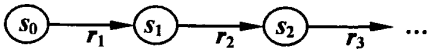
基金项目: 国家 863 计划资助项目(2001AA422200); 国家自然科学基金资助项目(60275024).
收稿日期: 2003 - 05 - 18

2 强化学习(Reinforcement learning)

强化学习的主要思想是"与环境交互(Interaction with Environment)"和"试错(Trial-and-error)". 下面介绍强化学习的标准模型[4,5].

2.1 基本模型

图 1 是强化学习的标准模型,主体与环境的交互接口包括行动(Action)、回报(Reward)和状态(State). 交互过程可以表述为如下模式:



这里的回报和指导学习(Supervised Learning)中的教师是不同的. 在指导学习中,教师要给出很多例子,告诉主体什么情况下,执行什么行动效果最好;在强化学习中,回报只告诉主体当前行动的执行效果,主体要在与环境交互过程中,不断测试每个行动的效果,在长时间的收集回报后,判断出每个行动的长远回报,完成主体的学习.

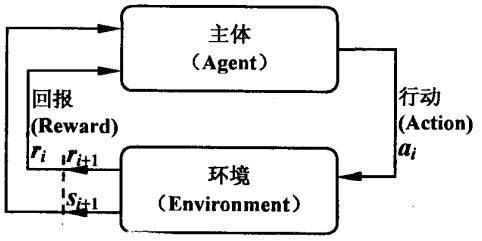


图 1 强化学习的标准模型

Fig. 1 The reinforcement learning framework

图 1 的强化学习模型事实上刻画了一类问题,只要一个问题可以描述成强化学习模型,那么解决这个问题的所有方法,都称为强化学习方法. 这样的方法有很多,下面先讨论强化学习中的几个关键概念:延迟回报(Delayed Reward)、探索和利用的权衡(Tradeoff between Exploration and Exploitation)及与未知环境交互.

2.2 延迟回报

主体为了完成给定的任务,必须知道每个行动的长远回报,而不仅仅是即时回报. 而长远回报必须经过一定时间的延迟之后才可以获得. 与环境交互过程中,t 时刻的延迟回报 R_t 可以用(1)式表示.

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots + \gamma^T r_{t+T+1} = \sum_{k=0}^{T} \gamma^k r_{t+k+1} \quad (1)$$

其中,$0 \leqslant \gamma < 1$ 是折扣率;对于有终任务(Episodic Tasks),T 是从当前开始到完成任务所需要的时间,对于持续任务(Continual Tasks),$T = \infty$.

如果环境的状态转换只与当前的状态和行动有关,与以前的状态和行动无关,即是一个马尔可夫过程(Markov Decision Process),就可以用(2)式表示每个状态的最佳长远回报,用(3)式表示每个状态-行动对的最佳长远回报:

$$V^*(s) = \max_{\pi} V^{\pi}(s) = \max_{\pi} E_{\pi}\{R_t \mid s_t = s\} = \ldots = \max_{a} \sum_{s'} P^a_{ss'}[\Re^a_{ss'} + \gamma V^*(s')] \quad (2)$$

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a) = \max_{\pi} E_{\pi}\{R_t \mid s_t = s, a_t = a\} = \ldots = \sum_{s'} P^a_{ss'}[\Re^a_{ss'} + \gamma \max_{a'} Q^*(s',a')] \quad (3)$$

其中 π 是选择行动的策略,P 是状态转换概率,$\Re$ 是即时回报的期望值. 有了这样的递归公式来描述问题,就可以用一定的方法迭代逼近最佳长远回报.

虽然很多问题不能用马尔可夫过程来描述,但以上的思想和方法仍然适用.

2.3 权衡探索和利用

假定主体在与环境交互的学习过程中,发现某些动作的效果较好,那么主体在下一次决策中该选择什么动作呢? 一种考虑是充分利用现有的知识,仍然选择当前认为最好的动作. 这样做有一个缺点:也许还有更好的动作没有发现. 反之,如果主体每次都测试新的动作,将导致有学习没进步,显然也不是我们愿意看到的. 所以,主体应该在利用现有知识和探索新的效果更好的动作之间做出适当的权衡.

权衡的策略有很多,简单的策略是每次以概率 ε 随机选择动作,以概率 $1-\varepsilon$ 选择当前认为最好的动作,该方法称为 ε-贪心法. ε-贪心法的缺点是在探索新的动作时,各个动作的选择概率是一样的,这样选择最坏行动的概率也很大.

还有一种常用的方法是以一定的概率选择各个动作,每个动作的选择概率和当前主体对它的评价值有关,一般以(4)式作为行动选择的概率:

$$p(a \mid s) = \frac{e^{Q(s,a)/T}}{\sum_{a'} e^{Q(s,a')/T}} \quad (4)$$

其中 $T > 0$ 是调控参数,T 越大,各个动作的选择概率越接近,$T \to 0$,则相当于贪心法.

以上两个选择动作的策略都只需要确定一个参数,一般来说,T 比 ε 更难确定一些:

2.4 与未知环境交互

在强化学习的很多问题中，主体事先不知道环境是什么样的，只有在与环境的交互过程中才知道. 如果出现下面的两种情况，还需要一定的方法来进行处理.

第一种情况是，环境的状态空间非常大，甚至是连续的，主体没有足够多的空间和时间来访问这些状态，这就要求使用一定的方法来泛化(Generalization)这些状态空间. 有时候主体的行动空间可能也很大，或是连续的，也要有一定的泛化方法. 另一种情况是，主体只能感知环境的部分信息，这就需要用部分马尔可夫过程(Partially Observable MDP)来描述问题，并且需要主体记录历史信息，即以前观察到的环境信息和执行过的行动等.

3 个人技术学习(Skills learning)

仿真机器人足球是一个动态、不确定、实时的多主体系统，它的状态空间是连续的，球员的动作空间也是连续的，且每个球员只能感知到部分的球场信息. 对于如此复杂的环境，直接应用强化学习方法非常困难. 如果把问题分解成一些子问题，如第一节中所述，可能会取得很好的效果. 本节介绍当前在个人技术层使用强化学习比较成功的一些例子，并做必要的分析.

个人技术是一个球员最基本的技能，如追球、拦球、传球、带球、射门、盯人等. 这些个人技能的任务都比较简单. 一个简单任务所需要的环境信息相对较少，球员的行动空间也大大减小.

3.1 动态规划

动态规划(Dynamic Programming)是强化学习中最常用方法之一. 由(2)式得到(5)式的迭代公式

$$V_{t+1}(s) \leftarrow \max_a \sum_{s'} P^a_{ss'}[\Re^a_{ss'} + \gamma V_t(s')] \quad (5)$$

$t \to \infty$ 时，$V_t(s) \to V^*(s)$，重复对每个状态 s 进行迭代，直到 $|\Delta V|$ 小于一个小正数，得到对每个状态长远回报的评价.

可以看出，动态规划需要事先知道状态转换概率和即时回报的期望值，即需要有环境的一个完整模型，可以离线进行学习.

很多问题事先很难给出环境的完整模型，仿真机器人足球就是这样一个问题. 文[6]用实时动态规划(Real-time DP)[7]来学习球员的个人技术. 实时动态规划不需要事先给出环境的模型，而是在真实的环境中不断测试，得到环境的模型. 主体首先根据当前对状态长远回报的评价选择下一步的动作，然后根据回报对以前访问过的状态 S^+ 进行迭代.

因为状态空间是连续的，文[6]在进行学习时，用了前馈神经网络(Feedforward Neural Network)对状态进行泛化[8]，网络的输入单元是环境的状态 s，网络的输出是对该状态的评价 $V(s)$. 在每一步迭代时，用(6)式得到输出单元的误差，再用前馈神经网络的反传(Back-propagation)算法修改网络的权重.

$$Error = \sum_{s_t \in S^+} (V(s_t) - (r_{s_t s_{t+1}} + V(s_{t+1}))^2) \quad (6)$$

仿真机器人足球的行动空间也是连续的，一般用量化(Quantization)的方法来进行泛化.

3.2 时序差分

时序差分(Temporal Difference)也是强化学习最常用的方法之一. 时序差分不需要环境的模型，每一步都根据即时回报，及时修改对长远回报的评价. 根据不同的修改方法，时序差分又分为很多不同的种类，文[9]和[10]用了时序差分中的 Q-Learning 方法来学习球员的个人技术.

Q-Learning 的迭代公式如(7)式所示：

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)] \quad (7)$$

其中 α 是学习速率. 如果 α 取值合适，在各种状态下，每个动作都被选中无限次，则 $Q(s,a)$ 以概率 1 收敛于最佳长远回报 $Q^*(s,a)$.

状态和行动空间仍然需要泛化，下面用文[9]中学习带球的例子加以说明. 用三个参量表示环境状态：对方球门方向、对手距离(离散为三个区间)和对手方向(窄视野范围内离散为四个区间，其它离散为两个区间)，用 5 个带球方向作为行动空间(假定球员对一个方向的带球已经学好)：-90°、-45°、0°、45°和 90°. 这样不仅可以把球带向对方球门，还可以避让对手. 如果对手把球得到，球员得到回报 $D - 0.01T$，其中 D 是到球门的距离，T 是总共花费的时间；带球中间每一步得到回报 $\Delta d - 0.01\Delta t$，其中 Δd 是从上一状态到现在带球带过的距离，Δt 是从上一状态到现在经过的时间.

4 战术学习(Tactics learning)

足球是一项集体运动，每个球员的技术再好，没有合作也是不行的. 用强化学习来学习合作也是很困难的，因为要考虑更多的因素. 如果个人技术已经做得比较好(通过学习或其它的方法)，在学习几个队友之间的战术配合时，用半马尔可夫过程(Semi-

Markov Decision Process)来描述问题,即直接把个人技术作为主体的行动空间,会大大简化战术的学习.下面我们就介绍一些在战术层成功使用强化学习的一些例子.

4.1 蒙特卡洛法

蒙特卡洛法(Monte Carlo methods)也是强化学习中最常用的方法之一,主要用于有终任务.蒙特卡洛法和时序差分一样,不需要环境的模型;和时序差分不同的是,每次要先根据策略,一步步地产生动作,直到任务完成,再对经过的状态或行动的评价进行统计更新.更新方法如下:

(1)用与当前长远回报评价有关的策略,不断产生动作直到一个任务完成;

(2)对任务中首次出现的每对(s, a):

①计算出在本次任务中的长远回报$R(s, a)$;

②把$R(s, a)$与以前各次任务中(s, a)的长远回报一起求平均,作为新的$Q(s, a)$;

(3)转(1).

文[11]提出了 TPOT-RL,结合了蒙特卡洛法和时序差分,来学习多个队友之间的传球和射门.因为它比较复杂,这里不再详述.

4.2 λ-时序差分

在第一节介绍了时序差分的 Q-Learning 方法,下面介绍时序差分的另一个方法:Sarsa,它的迭代公式如下:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)] \quad (8)$$

Sarsa 与 Q-Learning 的不同之处在于:Q-Learning 选择下一个状态中最大的行动评价作为更新依据,Sarsa 则根据下一个状态中实际选择的行动来更新评价.

时序差分的每一步都要对评价进行更新,如果不想每步都立即更新评价,而是过一定的时间步才更新评价,则可以引入一个参数 λ 来调整间隔步数,称为 λ-时序差分(TD(λ)).如果 $\lambda = 0$,则是一般的时序更新,如果 $\lambda = 1$,则相当于蒙特卡洛法.Sarsa(λ)更新评价的方法如下:

(1)对当前的状态s_t和选择的行动a_t,观察得到的回报r_{t+1}和下一步状态s_{t+1}:

①根据策略选择状态s_{t+1}下的行动a_{t+1};

② $\delta \leftarrow r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)$;

③ $e(s_t, a_t) \leftarrow e(s_t, a_t) + 1$;

(2)对所有的(s, a)对:

① $Q(s, a) \leftarrow Q(s, a) + \alpha\delta e(s, a)$;

② $e(s, a) \leftarrow \gamma\lambda e(s, a)$;

(3) $s_t \leftarrow s_{t+1}$, $a_t \leftarrow a_{t+1}$;

(4)如果任务没有完成,转(1);如果任务完成,设置初始(s_t, a_t)对,转(1).

文[12]用 Sarsa(λ)和 Tile 泛化来学习"3 对 2 控球(Keepaway)".文[13]用和 Sarsa 类似的桶队列(Bucket Brigade)算法来学习门前战术(对方球门前的决策:射门、传球或带球).除此之外,战术学习中使用比较多的方法仍然是第一节介绍的 Q-Learning.文[14]用 Q-Learning 和第一节介绍过的神经网络泛化来学习球员如何选择个人技术,文[15]用 Q-Learning 来学习"2 过 1",文[16]用 Q-Learning 来学习"2 过 1"和"2 过 2",文[17]用 Q-Learning 和 Coarse 泛化来学习"3 对 2 控球",文[18]用 Q-Learning 和 Kanerva 泛化来学习"3 对 2 控球".文[10]提出用多主体的 Q-Learning 和贝叶斯网络来学习局部战术配合和全局战术决策,但没有实现.

4.3 泛化

如果环境的状态空间非常大,甚至是连续的,就象仿真机器人足球一样,主体在进行强化学习时,没有足够多的空间和时间来访问这些状态,这就要求对状态空间进行泛化(Generalization).泛化方法在其它领域已经进行了充分的研究和应用,如机器学习、人工神经网络、模式识别和曲线拟合等.理论上说,这些领域内使用的泛化方法都可以应用到强化学习中,只要把强化学习和这些泛化方法很好地结合在一起.第三节中介绍的用神经网络进行泛化就是这样一个例子,一般称之为神经强化学习(Neural RL).下面再介绍一些上面提到的几个泛化方法.

Coarse 泛化是把一个多维连续空间转化为一组真假量,称为环境的二进制特征(Binary Feature).图 2 是一个二维空间 Coarse 泛化的例子,每个圆圈(可以是别的形状)是一个真假量,如果当前状态点在某个圆圈内,则对应的真假量为真,否则为假.

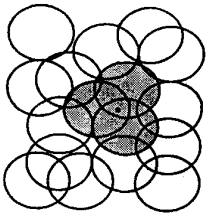


图 2 Coarse 泛化

Fig. 2 Coarse coding

Tile 泛化是 Coarse 泛化的一个特殊形式，它的一组真假量可以全部覆盖状态空间，并且每次只有一个真假量为真. 在二维情况下，可以用网格（可以是别的形状）来划分空间，每个格子是一个真假量，如图 3 所示. 如果对一个多维连续空间只用一次 Tile 泛化，则等同于量化（Quantization）. 为了使泛化更加精确，可以对状态空间进行多个 Tile 泛化，即用多组 Tile 真假量来泛化多维连续状态空间，如图 4 是二维情况下进行两个 Tile 泛化（可以采取两种不同的形状）.

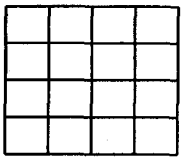


图 3　Tile 泛化

Fig. 3　Tile coding

如果是一个高维空间，比如上百维，Tile 泛化就很难实现. 但多数情况下，主体可能只会在其中很小的空间内活动. Kanerva 泛化根据主体的实际活动情况从高维状态空间中抽取出一些特征来表示状态，把这些特征都表示为二进制特征，比如用上面的 Tile 泛化. 然后 Kanerva 随机产生一些原型状态（Prototype States），对当前主体的二进制特征状态，计算它与各个原型状态的海明距离（Hamming Distance），距离最小的就认为和原型状态是一个状态.

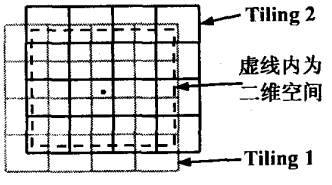


图 4　多个 Tile 泛化

Fig. 4　Multiple grid-tilings

文[18]在学习“3 对 2 控球”时，共抽取了 13 个特征，包括各个球员之间的距离、到控球中心的距离及最近队友、最近对手和控球队员形成的角度等. 每个特征用 Tile 泛化为一个三位二进制数，只有一位为真，数字越大，表示距离越远或角度越大. 这样就构成关于环境的 39 位二进制特征. 然后产生了 5000 个原型状态（每个原型状态是 39 位），每个输入状态都用海明距离归为原型状态中的一个. 这样，就把高维空间泛化为只有 5000 个状态的离散空间. 文[19]显示这样随机的状态表示可能比其它一些方法要好（比如最紧邻离散等），并且只要增加原型状态数目，就可以提高性能.

5　结论和展望（Conclusion）

强化学习的主要学习方法在仿真机器人足球中都有了很好的应用，而难点问题之一则是环境状态空间过大，如何把各种泛化方法和强化学习的各种方法很好地结合在一起，成为应用强化学习的首要考虑问题之一. 当前使用最多的泛化方法是简单的把连续量量化为离散量，使用比较成功的是文[6]中的神经网络泛化，其它效果比较好的是 Coarse 和 Kanerca 等线性泛化方法. 由于神经网络具有很好的泛化效果，神经强化学习将成为以后在仿真机器人足球中应用的主要方法之一，如神经网络和常用的 Q-Learning 结合而成的神经 Q-Learning（Neural Q-Learning）等[21].

关于仿真机器人足球环境的另一个难点问题是主体只能感知部分环境信息，文[10]提出用贝叶斯网络来描述状态，同时解决部分环境信息和环境状态空间过大问题，也将成为以后仿真机器人足球中应用的主要方法之一.

另外，文[10]还提出了多主体的 Q-Learning（Multi-agent Q-Learning）学习方法，这不仅会成为仿真机器人足球中应用强化学习的新方法，也对强化学习本身提出了新的研究方向.

参考文献　（References）

[1] Kitano H, Tambe M, Stone P. The RoboCup synthetic agent challenge 97[A]. Proceedings of International Joint Conference on Artificial Intelligence (IJCAI'97)[C]. 1997.

[2] Stone P. Layered learning in multi-agent systems: A winning approach to robotic soccer[A]. MIT Press[C]. 2000.

[3] Song Z W, Chen X P, *et al.* Layered decision-making and planning in shaoling team[A]. Proceedings of the 3rd World Congress on Intelligent Control and Automation[C]. 2000. 179 - 183.

[4] Sutton R S, Barto A G. Reinforment Learning: An Introduction[M]. MIT Press, 1998.

[5] Kaelbing L P, Littman M L, Moore A W. Reinforcement learning: a survey[J]. Journal of Artifial Intelligence Research, 1996, 4:237 - 285.

[6] Riedmiller M, Merke A, Meier D, *et al.* Karlsruhe Brainstormers - A Reinforcement Learning Approach to Robotic Soccer. RoboCup-2000:

Robot Soccer World Cup IV[M]. Springer, 2000.

[7] Barto A G, Bradtke S J, Singh S P. Learning to act using real-time dynamic programming[J]. Artifical Intelligence, 1995, (72):81 - 138.

[8] Riedmiller M. Concepts and facilities of a neural reinforcement learning in cooperative multi-agent systems[J]. Journal of Neural Computing and Application, 2000, 8:323 - 338.

[9] Arseneau S, Sun W, Zhao C P, *et al*. Inter-layer learning towards emergent cooperative behavior[A]. American Association for Artificial Intelligence (AAAI 2000)[C]. Austin, Texas:, 2000.

[10] Maes S, Tuyls K, Manderick B. Reinforcement learning in large state spaces simulated robotic soccer as a tested[A]. Pre-Proceedings of RoboCup 2002: Robot Soccer World Cup VI[C]. 2002.

[11] Stone P, Veloso M. Team-partitioned, Opaque-transition Reinforcement Learning. RoboCup-98: Robot Soccer world Cup II[M]. Springer, 1998.

[12] Stone P, Sutton R S. Scaling reinforcement learning toward RoboCup soccer[A]. Proceedings of the 18th International Conferences on Machine Learning[C]. 2001.

[13] Lind J, Jung C G, Gerber C. Adaptively and learning in intelligent real-time systems[A]. Proceedings of the Third International Conference on Autonomous Agents[C]. 1999.

[14] Riedmiller M, Buck S, Dilger S, *et al*. Karlruhe brainstormers-design principles[A]. RoboCup-99 Team Descriptions, Simulation League, Team Karlsruhe Brainstormers[M]. 1999. 59 - 63.

[15] Hu H S, Kostiadis K, Liu Z Y. Coordination and learning in a team of mobile robots[A]. Proceedings IASTED Robotics & Applications Conference[C]. California: 1999.

[16] Kostiadis K, Hu H S. Reinforcement learning and co-operation in a simulated multi-agent systems [A]. Proceedings of IEEE/RJS IROS'99[C]. Korea: 1999.

[17] Stone P, Sutton R S, Singh S. Reinforcement learning for 3 vs. 2 Keepaway[A]. RoboCup 2000: Robot Soccer World Cup IV[M]. Springer, 2000. 249 - 258.

[18] Kostiadis K, Hu H S. KaBaGe-RL: Kanerva based generalization and reinforcement learning for possession football[A]. Proceedings IEEE/RSJ International Conference on Intelligent Robots & Systems (IROS 2001)[C]. Hawaii: 2001.

[19] Sutton R S, Whitehead S D. Online learning with random representations[A], Proceedings of the 10th International Conference on Machine Learning[C]. Morgan Kaufmann: 1993.314 - 321.

[20] Ohta M. Gemini-learning cooperative behaviors without communicating[A]. RoboCup-99 Team Descriptions. Simulation League, Team Gemini[M]. 1999. 36 - 39.

[21] Hagen S, Kröse B. Q-learning for systems with continuous state and action spaces[A]. Proceedings of the 10th Belgian-Dutch Conference on Machine Learning[C]. 2000.

作者简介:

宋志伟 (1978-),男,博士生,研究领域:多主体系统,机器学习等.

陈小平 (1955-),男,教授,博士生导师,研究领域:智能体理论,机器学习,多机器人系统等.

仿真机器人足球中的强化学习

作者： 宋志伟， 陈小平
作者单位： 中国科学技术大学计算机科学技术系,安徽,合肥,230027
刊名： 机器人 ISTIC EI PKU
英文刊名： ROBOT
年，卷(期)： 2003，25(z1)
引用次数： 0次

参考文献(4条)

1. Riedmiller M. Merke A. Meier D Karlsruhe Brainstormers - A Reinforcement Learning Approach to Robotic Soccer. RoboCup-2000:Robot Soccer World Cup Ⅳ 2000

2. Maes S. Tuyls K. Manderick B Reinforcement learning in large state spaces simulated robotic soccer as a tested 2002

3. Stone P. Veloso M Team-partitioned, Opaque-transition Reinforcement Learning. RoboCup-98: Robot Soccer world Cup Ⅱ 1998

4. Stone P. Sutton R S. Sinsh S Reinforcement learning for 3 vs. 2 Keepaway 2000

相似文献(5条)

1. 期刊论文 殷翔. 黄展翔 强化学习在仿真机器人足球踢球动作中的应用 -苏州大学学报(工科版)2002,22(4)
机器人足球中的环境复杂性是对强化学习方法的一大挑战.本文介绍了在科大蓝鹰仿真机器人足球队采用的强化学习技术,将强化学习方法与启发式搜索策略结合后,针对底层动作中的踢球问题提出了一种实际可行的解决方案.

2. 学位论文 Nadeem Iqbal RoboCup仿真机器人足球多代理系统的机器学习研究与应用（英） 2002
该文对自主智能体组成的多代理系统,在实时、噪音、合作及对环境中的机器学习和协调控制进行了研究,将强化学习应用在robocup仿真机器人足球中,改善了智能体在射门、截球、运球的能力,并通过仿真实验得到证实.

3. 学位论文 张晓红 RoboCup仿真环境下Agent机器学习策略的研究 2007
RoboCup(机器人世界杯足球赛)是国际上规模最大、影响最为广泛的机器人足球赛事，它是人工智能和机器人研究的一种集中体现。在RoboCup的仿真环境下，机器学习的很多方法都可以得到检验。作为人工智能和机器人学新的标准问题，机器人足球受到越来越广泛的关注。 本文使用仿真机器人足球作为研究机器学习算法应用的载体，基于Agent和机器学习的理论，围绕Agent个体的基本动作及Agent个体高级动作的决策问题展开研究，主要研究内容和成果如下： 1．将强化学习的方法引入到Agent踢球动作中，实现了把球加速到指定的出球速度的目的；在简化条件下采用BP神经网络的方法拟合Agent踢球力量和速度的关系，以便应用到简单的踢球动作的选择中。 2．改进现有的神经网络的截球方法，构造出一种新的基于BP神经网络的截球动作模型，并将该模型直接应用到传球的判断中，实现了传球路线的选择：将RBF神经网络引入到射门模型中，根据强化学习和BP神经网络结合的思想提出了一种新的基于前向神经网络的Q学习的算法，较好的实现了射门模型。

4. 期刊论文 蔡庆生. 张波. CAI Qing-Sheng. ZHANG Bo 一种基于Agent团队的强化学习模型与应用研究 -计算机研究与发展2000,37(9)
多Agent学习是近年来受到较多关注的研究方向.以单Agent强化学习Q-learning算法为基础 ,提出了一种基于Agent团队的强化学习模型,这个模型的最大特点是引入主导Agent作为团队学习的主角,并通过主导Agent的角色变换实现整个团队的学习.结合仿真机器人足球领域 ,设计了具体的应用模型,在几个方面对Q-learning进行了扩充,并进行了实验.在仿真机器人足球领域的成功应用表明了这个模型的有效性.

5. 学位论文 方宝富 MAS结构和协作机制研究及其在Robocup中的应用 2003
分布式人工智能是人工智能的一个分支,已经成为当前的研究热点.而对分布式人工智能最为有效的求解方式是基于Agent技术的建模.Agent建模主要是设计合适的Agent结构,通过学习或其他方法得出Agent的智能,最后使用一定的协调方法和规划达到Agent之间的协作.该文在基于Agent、MAS的理论基础上设计了一个实际能够使用的MAS系统—Robocup仿真球队.该MAS系统充分考虑了系统的实时性和噪音,系统设计的每个Agent具有合理的结构和相当的智能性并且能够根据环境做出比较协调的协作动作.在该文中,首先介绍了典型的Agent结构和MAS模型和仿真机器人足球的一些主要模型;设计了一个分层的Agent结构—HfutAgent,通过机器学习算法实现了Agent的个体智能;最后结合足球领域专家的知识实现了Agent间的协作,其中使用了Robocup中一个典型的协作方法-SBSP,设计了一个通过强化学习的方法来达到Agent之间的局部协作,把基于效用的对策论方法引入了HfutTeam的进攻体系和防守体系中.