模糊 Q学习的足球机器人双层协作模型

曹卫华,徐凌云,吴 敏
(中南大学 信息科学与工程学院,湖南 长沙 410083)

摘 要:针对传统的足球机器人 3层决策模型存在决策不连贯的问题和缺乏适应性与学习能力的缺点,提出了一种基于模糊 Q学习的足球机器人双层协作模型. 该模型使协调决策和机器人运动成为 2个功能独立的层次,使群体意图到个体行为的过度变为一个直接的过程,并在协调层通过采用 Q学习算法在线学习不同状态下的最优策略,增强了决策系统的适应性和学习能力. 在 Q学习中通过把状态繁多的系统状态映射为为数不多的模糊状态,大大减少了状态空间的大小,避免了传统 Q学习在状态空间和动作空间较大的情况下收敛速度慢,甚至不能收敛的缺点,提高了 Q学习算法的收敛速度. 最后,通过在足球机器人 SimuroSot仿真比赛平台上进行实验,验证了双层协作模型的有效性.

关键词:足球机器人;双层决策模型;基于行为的控制系统;Q学习
中图分类号: TP18 **文献标识码**: A **文章编号**: 1673-4785(2008)03-0234-05

A double-layer decision-making model based on fuzzy Q-learning for robot soccer

CAO Wei-hua, XU Ling-yun, WU Min
(School of Information Science and Engineering, Central South University, Changsha 410083, China)

Abstract: With the conventional triple-layer decision-making model of soccer robots, decisions are sometimes inconsistent, leading to weaknesses in adaptability and self-learning ability. A double-layer cooperation model for a robot soccer system based on fuzzy Q-Learning is presented to solve these issues. This model divides cooperative decisions and robot movement into two layers with their own independent functions, so that the transition from group strategy to individual behavior becomes a direct process. To enhance the adaptability and self-learning capabilities of the decision-making system, the Q-learning algorithm was used in the cooperation layer to learn the optimal strategy for various conditions. To speed up the convergence of Q-learning and decrease the size of the state space, the numerous system states were mapped to seven fuzzy states in Q-learning. This avoids problems with Q-learning's slow converging rate when the size of the state space is large. This model was verified on the SimuroSot Robot Soccer Game platform.

Keywords: robot soccer; double-layer decision-making model; behavior-based control system; Q-learning

近年来,对于多智能体的研究已经成为人工智能研究领域的重要方向和热点,其中多智能体协作模型的研究最为瞩目. 很多学者提出使用机器学习的方法来实现多智能体之间的协作和协调,例如遗传算法、神经网络和强化学习等.

20世纪 80年代末,随着分布式人工智能的发展,多智能体技术逐渐被应用到各种多机器人系统中. 足球机器人是一个典型的多智能体系统,在 FIRA系列足球机器人系统中一般包括 4个子系统:视觉子系统、决策子系统、通信子系统和机器人车体子系统,它具有比赛环境复杂、难以建立准确数学模型等特点. 其中,决策子系统是整个系统的关键,它负责接收经视觉子系统处理后的场地信息,通过设计的决策算法实现机器人的协作和控制,因此决策子系统设计的好坏直接关系到整个系统的性能[1]. 一种广泛使用的方法采用基于专家经验和规划的 3层结构模型来设计决策子系统[2-3]. 3层结构的模型虽然在逻辑结构上非常清晰,但是存在决策不连贯

收稿日期: 2007-11-15.
基金项目:湖南省自然科学基金资助项目 (06JJ50144).
通讯作者:吴 敏. E-mail: min@csu.edu.cn.

的缺点[4].文献[5]提出了一种基于行为的双层决策模型,能够解决 3层结构的缺点,但它也是采用基于专家经验的方法来实现机器人间的协作,因此决策系统缺乏适应性和自学习能力.

通过在协调层中引入 Q学习算法,针对 Q学习算法在收敛性方面的不足,提出一种基于状态空间模糊化 Q学习的足球机器人双层决策模型,在一定程度上提高了决策系统的适应性和自学习能力.

1　双层决策模型结构与模糊 Q学习

决策子系统通过视觉子系统获得比赛场地上的各种信息,通过分析这些信息做出决策,并最终产生各个机器人的左右轮速发送给通信子系统.双层决策模型把整个决策子系统分为 2个层次:协调层和运动控制层,总体结构图如图 1所示.

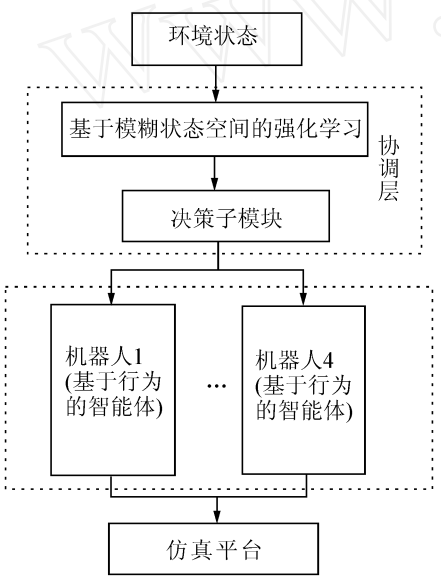


图 1　双层决策模型总体结构图

Fig 1　Structure of the double-layer decision-making model

协调层用于产生机器人的群体意图.它把整个决策意图划分为 4个子决策模块:积极进攻模块、保守进攻模块、禁区防守模块、重叠防守模块,然后通过分析当时比赛环境和对手的策略,采用基于模糊状态空间的 Q学习方法来选择最优决策子模块,使得己方策略能够最大限度压制对手决策,从而能在比赛中获得优势;最后把决策意图发送给运动控制层.

在运动控制层中,机器人接收上层决策传送过来的机器人全局目标描述并把这个全局目标描述转化为自己局部视觉描述,然后针对每个机器人采用基于行为方法来设计其行动方式,使得每个机器人成为一个有自主路径规划等决策能力的单独智能体[5].本文主要针对强化学习在协调层中研究,以提高双层决策模型的适应性和自学习能力.

Q学习是由 Watkins于 1989年提出来的一种模型无关的强化学习算法[6-9].单步 Q学习算法的基本形式如式(1)所示:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha(r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)). \quad (1)$$

式中:γ为折扣率,α为学习率(或学习步长).Agent在 s_t状态下,根据策略选择函数确定动作 a_t,得到奖赏值和训练例 $(s_t, a_t, s_{t+1}, r_{t+1})$;然后根据此奖赏值,依据式(1)修改 Q值.当 Agent访问到目标状态,算法终止一次迭代循环.研究表明,当 α满足一定条件时,Q学习算法必然收敛在最优解[10].

虽然 Q学习在单个 Agent领域应用非常成熟,也较多地应用于多智能体领域,但是运用于足球机器人决策还存在以下问题:Q学习算法在状态空间和动作空间较大的情况下,学习效率低,收敛速度慢.而足球机器人比赛的状态复杂繁多.

针对此问题,通过分析足球机器人的行为特征及双层决策模型的特点,仅在上层进行决策的 Q学习.这样,机器人的基本行为动作集中体现在下层,采用基于行为的控制方式,缩小了 Q学习的动作空间;在上层进行的 Q学习中,通过引入状态模糊化的思想,把状态繁多的精确状态映射到 7个模糊状态,进一步缩小 Q学习的状态空间.这种双层决策模型大大降低了 Q学习原始状态空间的维数.

2　基于模糊 Q学习的协调层设计

足球机器人系统是一个多智能体系统,它的环境复杂多变.因此,在协调层算法设计中需要解决状态空间划分、策略库设计、奖励函数设计等问题,它们设计的好坏直接关系到整个系统的收敛性问题.下文将具体阐述这些问题的设计方法.

2 1　协调层模型

通过以上分析,设计了如图 2所示的基于强化学习的协调层模型.其中,s为仿真环境当时的状态输出;r为环境的奖赏值;s'为比赛进行后的环境的改变;i为感知器通过对原始环境的处理后,得出的各个机器人的信息、球的信息和对手的策略等.

首先,感知器通过分析环境 s得出场上形势和对手的策略 i;然后,策略选择器选择相应的策略 n下发到运动控制层,并由运动控制层把具体的策略作用到仿真环境上;最后,仿真环境状态变化为 s',并产生相应的强化信号 r给学习器,学习器通过 r来更新策略库.这样,通过很多周期的反复迭代,协

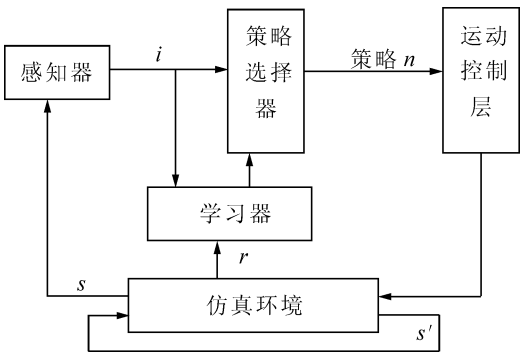


图 2 协调层结构图

Fig 2 Frame of the cooperation layer

调层就可以学习到在具体情况下的最优策略.

2.2 状态空间的划分与状态映射函数

由于足球机器人比赛环境复杂,形势瞬息万变. 如果状态集合设计得不合理,会使得 Q学习状态空间非常庞大,这将直接影响到强化学习算法的收敛速度. 因此设计合适的状态空间十分关键. 针对这种情况,本文采用模糊化的方法把实际状态转化为模糊状态,从而一定程度上降低了状态空间的大小. 以我方从左向右进攻为例,其结构图 3所示. 将整个场地划分为 11个区间、7种状态,这样,状态空间即可表示为 $S = \{$底线有利、边路有利、中路有利、中路相持、边路威胁、中路威胁、底线威胁$\}$. 以我方从左向右进攻为例,其状态划分如图 3所示.

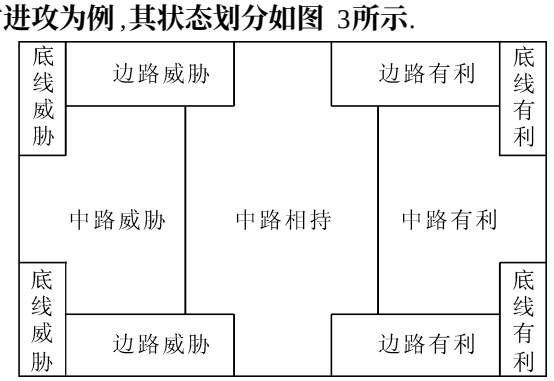


图 3 状态划分图

Fig 3 Schematic diagram for state division

在状态划分中,考虑 3个主要的环境因素: 1) 球在场上的坐标位置 $p(x, y)$,依然将其模糊化:根据图 3,将球的位置状态模糊为 7种状态 p_k, $k \in [1, 7]$; 2) 球的速度 v及运动角度 θ; 3) 对方球员在我方球门附近的个数 n;. 因此设计了把 1)、2)、3)作为模糊模块的输入,输出为场上状态的映射函数,如式 (2).

$$s(p_k, v, \theta, n) = \omega \cdot p_k (v \cdot \tan\theta + n). \quad (2)$$

式中:ω为模糊因子.

2.3 学习算法设计

针对以上分析,采用极小极大 Q算法来设计协调层. 通过最小化对方决策的奖赏来选择使己方获得最大奖赏的策略. 其值函数如下

$$V(s) = \max_{d \in D} \min_{o \in O} \sum_{d \in D} Q_c(s, d, o). \quad (3)$$

协调层的 Q函数 $Q_c(s, d, o)$的更新规则为

$$Q_c(s, d, o) \leftarrow (1 - \alpha) Q_c(s, d, o) + \alpha (r_c(d, o, s) + \gamma V_c(s')). \quad (4)$$

式中: s为当前的环境状态, d为己方在状态 s下的策略, D为己方策略集合, o为对手在状态 s下的策略, O为对方策略集合, r_c为强化信号, s'为新的环境状态, α为学习率, γ为折扣因子.

具体学习过程如下

1) 观察当前状态 s.

2) 通过以下方法选择一个动作 d:

$$P(d_i \mid s) = \frac{\exp(Q(s, d_i)/T)}{\sum_{a_k \in A} \exp(Q(s, d_k)/T)}. \quad (5)$$

式中: T称作温度值. 式 (5)用来确定随机策略的随机度. 这种方法被称作 Boltzmann选择方法,也可采用 ε贪心策略等探测方法.

3) 观察新的状态 s'.

4) 从环境中获得即时回报 r.

5) 根据式 (4)对状态 s和动作 d相应的 Q值进行更新.

6) 如果新的状态 s满足结束条件,结束这次学习;否则 $s \rightarrow s'$,返回 2).

2.4 策略设计

策略设计分为己方策略 D设计和对方策略 O设计. 由于协调智能体的输出为决策子模块,这使得策略维数大大降低. 通过分析,本文设计了 4个决策子模块,即己方策略 $D = \{$积极进攻,保守进攻,重叠防守,禁区防守 $\}$. 由于采用的是极小极大 Q算法,所以对方策略 $O = \{$ 禁区防守,重叠防守,保守进攻,积极进攻 $\}$.

2.5 基于目标的奖赏函数设计

奖励函数是决定智能体执行特定动作后环境给它的强化信号. 传统的 Q学习中,如果决策目标达成,则得到正的奖励 m,反之则得到负的奖励 $-m$. 如式 (6)所示:

$$r_s = \begin{cases} m, & \text{己方进球;} \\ -m, & \text{对方进球, } m > 0; \\ 0, & \text{其他.} \end{cases} \quad (6)$$

但是很多情况下,动作执行的结果很难在一开始就看到. 例如,一个进球是由很多决策加在一起后执

行的结果.因此,采用单纯的传统设计方法,Q学习收敛速度将会很慢,很难达到要求.针对这种情况,本文设计了一系列基于目标的奖励函数设计方法.

在足球机器人中,最主要的目标是小球.所以衡量一个策略有效性的重要指标就是策略执行后球的位置、运动方向、速度是不是朝着策略的目标方向行进.将球在场上的位置做 x、y的坐标分解,可以看出,球在 x方向上与球门的距离关系更为重要.针对这种情况,本文设计了一种基于球在一个周期内在移动的水平距离 S_x与球在一个周期内能够移动的最大水平距离 S_{max}的比值的奖励分配方法.以球在球场上的位置为原点作局部坐标系,如图 4所示.

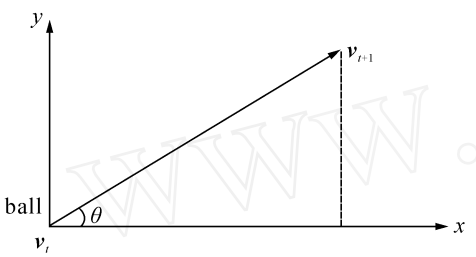


图 4　目标坐标分解图

Fig 4　Schematic diagram for target coordinate disassembling

图 4中,v_t和 v_{t+1}分别为球在 t时刻和 $t+1$时刻的速度,x轴为己方进攻方向(具有正负方向),θ为球的运动方向角.则球运动的距离 S_x为

$$S_x = \cos\theta((V_t + V_{t+1}) \cdot t/2). \quad (7)$$

假设机器人最大速度为 V_{max},那么在时间 t内,运行的最大水平距离为 $S_{max} = V_{max} \cdot t$

综上所述,本文设计的基于水平距离的奖赏函数如式(8)所示:

$$r_a = m \cdot (S_x / (V_{max} \cdot t)). \quad (8)$$

式中:m与式(6)中的 m相同,为最大奖励值,速度 V的正方向跟 x轴的正方向相同.

除了上述奖励方法,还设计了其他一些有效的奖励方法,并综合输出奖惩值 r_c:

1) 如果一个防守队员转化为进攻队员,则奖励,反之,如果一个进攻队员转化为防守队员,惩罚;

2) 将模糊化的场上状态空间分为 3种类型:有利状态(底线有利、中路有利、边路有利),中间状态(中场相持),被动状态(底线威胁、中路威胁、边路威胁).参考本周期的场上状态 s到下一周期的场上形势 s',如果从被动状态转化为中间状态或有利状态,则奖励.如果从有利状态转化为中间状态或被动状态,则惩罚.

最后,通过综合考虑 3类奖励信息,按照加权的方法得到了如式(9)所示的奖励函数:

$$r_c = \omega_a \cdot r_a + \omega_b \cdot r_b. \quad (9)$$

式中:ω_a、ω_b为对应的加权系数,且 $\omega_a, \omega_b \geq 0, \omega_a + \omega_b = 1$.

3　仿真实验

仿真实验是在足球机器人仿真平台 Robot Soccer v1.5a上进行的.

按照上述方法进行设计协调层,并采用文献[5]中基于行为的运动控制层来编写决策算法,在足球机器人仿真平台上进行了算法仿真.算法的参数设置为:折扣因子 $\gamma = 0.9$,学习率 α的初始值设为 0.8.仿真过程中,对手采用随机策略选择动作.

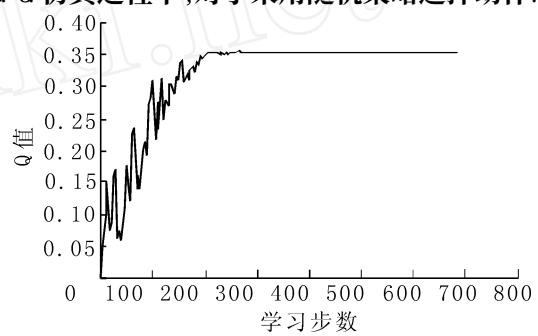


图 5　Q值曲线

Fig 5　Q-value curve

首先实验在经过 500步学习后对学习效果进行观察.图 5显示了在一次进攻中进攻机器人的积极进攻策略 Q值情况.从图中可以看出,随着学习步数地增加,Q值快速上升,最后收敛于一个稳定的值.所以,基于模糊 Q学习的决策算法收敛速度快,学习效率高,算法稳定.

然后对 500场比赛内每 20场比赛的平均净胜球数进行统计,分别采用基于模糊 Q学习的双层协作模型策略和传统的基于专家经验的策略进行比赛,并将结果进行分析对比,结果如图 6所示.其中,坐标 y表示净胜球数:

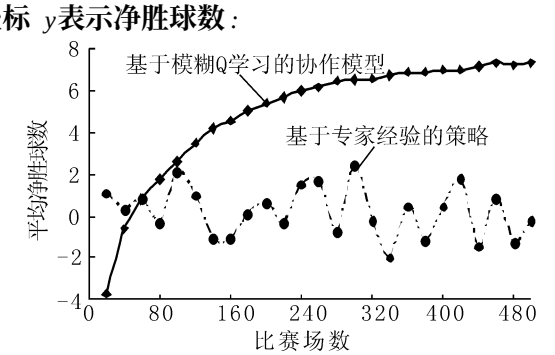


图 6　净胜球统计

Fig 6　Statistics for goal difference

净胜球数 = 我方进球数 - 对方进球数.

从图 6中可以看出,基于模糊 Q学习的双层协作模型的净胜球趋势线是上升的.在学习的开始阶段(大约第 50场比赛之前),赢的场数比输的场数要少.这表明系统还在探索学习,因此学习的效果并没有采用经验知识的策略好.但是趋势线上升明显,说明学习的方法逐渐显示出了它的作用.从第 50场开始,赢的场数比输的场数逐渐增多,平均净胜球个数成为正数.从第 120场开始,趋势线的上升开始逐渐变缓.这是由于在进行了多场比赛后,Q学习已经学习到大多数对方的策略.而基于传统专家经验策略的净胜球趋势线显示出了明显的随机性,净胜球数在 0左右震荡,没有明显的规律性.图 6中的趋势线表明基于 Q学习的策略是有效的,并且比赛的结果也是越来越好.基于该协作模型的机器人系统在 2006年全国机器人赛 Middle League SimuroSot项目中获得二等奖的好成绩.

4 结束语

提出足球机器人双层决策模型并在协调层中引入模糊 Q学习方法,它能够有效地解决传统决策模型中由于决策交错而引起的机器人运动不连贯的缺点,并提高了决策系统的适应性和自学习能力.在协调层中,针对机器人足球比赛的特点,通过采用对比赛场地分区和位置映射的方法,大大地降低了状态空间大小.另外,提出了一种基于水平距离的奖励函数设计方法,使奖励分配更加合理,这在一定程度上提高了 Q学习的收敛速度.

参考文献:

[1] ASADA M, KITANO H. The robocup challenge[J]. Robotics and Autonomous System, 1999, 29(1): 3-12.

[2] 赵逢达, 孔令富, 李贤善. 基于分层结构模型的机器人足球决策系统设计[J]. 哈尔滨工业大学学报, 2005, 37(7): 933-935.
ZHAO Fengda, KONG Lingfu, LI Xianshan. Design of robot soccer decision-making subsystem based on layered structure model[J]. Journal of Harbin Institute of Technology, 2005, 37(7): 933-935.

[3] 陆永忠, 柯文德. 足球机器人决策系统的设计与实现[J]. 计算机仿真, 2007, 24(9): 129-132.
LU Yongzhong, KE Wende. Design and implementation of decision system for soccer robot[J]. Computer Simulation, 2007, 24(9): 129-132.

[4] 刘云江, 韩光胜. 基于多智能体规划的机器人足球决策模型[J]. 哈尔滨工业大学学报, 2004, 36(7): 871-873.
LIU Yunjiang, HAN Guangsheng. Decision-making model for robot-soccer based on multi-Agent[J]. Journal of Harbin Institute of Teachnology, 2004, 36(7): 871-873.

[5] 曹卫华, 桂卫华, 吴 敏, 等. 一种基于行为的足球机器人双层决策模型[C]//哈尔滨: 2006中国控制会议论文集. 2006: 871-873.
CAO Weihua, GUI Weihua, WU Min, et al. A double-layer decision-making model based on behavior[C]// Proceedings of the 25th Chinese Control Conference (Ⅲ). Harbin: 2006: 871-873.

[6] 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 Rui, WU Min, PENG Jun, et al. A new Q learning algorithm for multi-Agent systems[J]. Acta Automatica Sinica, 2007, 33(4): 367-372.

[7] 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.
GAO Yang, CHEN Shifu, LU Xin. Research on reinforcement learning technology: a review[J]. Acta Automatica Sinica, 2004, 30(1): 86-100.

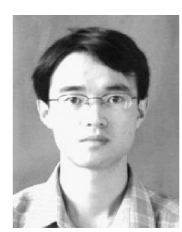
[8] WATKINS C H, DAYAN P. Technical note: Q-learning[J]. Machine Learning, 1992, 8(3-4): 279-292.

[9] TSITSIKLIS J N. Asynchronous stochastic approximation and Q-learning[J]. Machine Learning, 1994, 16(3): 185-202.

[10] LITTMAN L, SZEPESVARI C. A generalized reinforcement learning model: convergence and Applications[C]//Proc of the 13th International Conference on Machine Learning. Bari, Italy: Morgan Kanfmann, 1996: 310.

作者简介:

曹卫华,男,1972年生,副教授、博士,主要研究方向为机器人与智能系统技术和过程控制. 1996～1997年赴日本金泽大学留学一年.获省部级科技进步二等奖 2项、三等奖 2项.



徐凌云,男,1982年生,硕士研究生,主要研究方向为足球机器人系统与多智能体技术.

吴 敏,男,1963年生,博士生导师,主要研究方向为过程控制、鲁棒控制和智能系统.1989～1990年在日本东北大学进修;1996～1999年赴日本东京工业大学从事国际合作研究;2001～2002年得到英国皇家学会资助,为英国诺丁汉大学访问教授.1999年与中野道雄教授和佘锦华博士一起获国际自动控制联合会(IFAC)控制工程实践优秀论文奖.