多智能体强化学习在足球机器人中的研究与应用

刘春阳[1,2],谭应清[1],柳长安[1],马莹巍[1]
(1.华北电力大学控制与计算机工程学院,北京 102206;2.北京科技大学信息工程学院,北京 100083)

摘 要: 本文提出一种基于投票的多智能体强化学习方法,使球队在比赛中学会协作,自动适应环境,提高实时性和进球数.首先通过定义称为策略的联合行为,将协作问题转化为对策略的学习,简化问题的处理;然后对球场进行划分,以区域表示位置,有效减少了状态空间维数,加快了学习速度;接下来通过区分环境状态并只考虑协作状态,减小状态空间,进一步提高了学习速度;并使用投票的方式综合各个队员的决策,达到协作的目的.最后通过实验结果表明了该方法的正确性和有效性.

关键词: 强化学习; 机器人足球; 多智能体系统; 投票
中图分类号: TP242.6 **文献标识码:** A **文章编号:** 0372-2112 (2010) 08-1958-05

Application of Multi-Agent Reinforcement Learning in Robot Soccer

LIU Chun-yang[1,2], TAN Ying-qing[1], LIU Chang-an[1], MA Ying-wei[1]
(1. *School of Computer Science and Technology, North China Electric Power University, Beijing* 102206, *China*;
2. *University of Science and Technology Beijing, Beijing* 100083, *China*)

Abstract: A multi-agent reinforcement learning method based on voting to solve the collaboration problem of team members is presented. The method translates the collaboration problem into learning strategies by defining joint actions which called the strategies and then can simplify the problem. Through dividing of the playground, the location can be measured by a lot of numbered regions and then can effectively reduce the state-space dimensions to speed up the pace of learning. By distinguishing the environment states and taking the collaboration status into account, that causing the reduction of the state-action space, the learning speed can be further improved. Using a voting process that combines the decisions of the agents can realize the collaboration. At last, experimental results show the effectiveness and correctness of the method.

Key words: reinforcement learning; robot soccer; multi-agent system; vote

1 引言

近年来,多智能体系统 MAS 已成为人工智能领域中一个引人注目的分支[1].多智能体系统是由多个可计算的智能体(Agent)组成的集合,它能协调一组自主体的行为(知识、目标、方法和规划等),从而协同地动作和求解问题.机器人足球比赛是一个有趣且复杂的人工智能研究领域,参赛的足球机器人通过交互共同完成任务,涉及多智能体在动态环境中实时知识处理过程,融合了实时视觉系统、机器人控制、无线通讯、多机器人控制等多个领域的技术,被称为是"一个小平台上的技术战争"[2].它已经成为研究多智能体系统的一个标准实验平台.

强化学习是一种在线学习方法,它采用类似于人类思维中常常存在的一种"试错—修改"[3]的过程,借助反馈信息并以合适的算法强化好的行为和弱化差的行为,收敛为最优行为,适合学习者对环境了解甚少,或者动态环境下的问题.与监督学习技术通过正例、反例来告知采取何种行为不同,强化学习通过"试错"与环境交互获得策略的改进,其自学习和在线学习的特点使其成为机器学习研究的一个重要分支[4].

Q-学习是无需环境模型的一种强化学习形式,它提供 Agent 在 Markov 环境中利用经历的动作序列执行寻找最优动作的一种能力,并且不需要建立环境模型.马尔可夫决策过程(MDP, Markov Decision Processes)是强化学习的数学模型,因此,通常顺序型任务中的强化学习问题可以通过马尔可夫决策过程建模[5].用 $Q^*(s, a)$表示 Agent 在状态 s 采用动作 a 所获得的最大折扣奖赏.Agent 针对环境状态的最优策略为在每一环境状态选用 Q 值最大的行为,从环境因素对行为进行选择.

收稿日期:2008-12-18;修回日期:2010-01-08
基金项目:国家自然科学基金(No.60775058);华北电力大学青年教师科研基金项目(No.200721006)

$Q^*(s,a)$和最优策略$\pi^*(s,a)$求解如下:

$$Q^*(s,a) = R(s,a) + \gamma \sum_{s' \in S} T(s,a,s') \max_{<a'>} Q^*(s',a') \tag{1}$$

其中 γ 为折扣因子[6].

$$\pi^*(s,a) = \arg \max_{<a \in A>} Q^*(s,a) \tag{2}$$

2 多智能体强化学习

多 Agent 强化学习是在单 Agent 强化学习的基础上发展起来的,但目前大多数多 Agent 学习方法还仅限于将单 Agent 的学习方法简单移植到多 Agent 领域.单 Agent 学习考虑的是提高自己的问题求解能力,体现的是个体的智能和适应性;而多 Agent 学习更多考虑的是如何提高整体的自适应能力,体现的是群体智能和社会性.单 Agent 学习和多 Agent 学习由于学习目的和社会特性的不同,导致了双方在学习过程的模型和算法的不同.

单 Agent 强化学习一般都是以马尔可夫决策过程为模型化的依据,在此基础上使用值迭代或者策略迭代的方法求解[6].单 Agent 强化学习常常是独立学习,在学习过程中认为环境是固定的,并不考虑 Agent 的行为间的相互关系,学习方式是集中式学习,不涉及过多的通信和交互.多 Agent 强化学习本质上是一个非固定的动态过程,继续使用马尔可夫决策过程来模型化已不合适[7].多 Agent 强化学习是群体学习,每个 Agent 不仅需要自己学习,也和其他 Agent 共享信息并获得他们的相关知识,涉及 Agent 之间的交互和通信.系统中每个 Agent 的学习和行为都会导致整个系统的变化,因此 Agent 在学习过程中,行为的确定不仅和当前的状态有关,同时也与其他 Agent 的状态和所采取的行为有关.通常多 Agent 系统的强化学习应该是分布、并行和容错的[8].

在机器人足球这样的团体比赛中,队员之间需要相互协作并对复杂多变的赛场情况迅速作出反应,因此球队应该具有学习能力以适应新的环境.多个队员如何学习在复杂环境下协作,以最大化团队行为的期望收益是本文要解决的关键问题,也是学习的目的.机器人足球系统中的很多因素如球员数量的增加导致决策的指数级增长、赛场环境的部分可观察性、分布式环境下队员某些时候无交互的自主决策行为造成环境状态的不可预测性以及环境噪音干扰等问题给传统强化学习提出了严峻的挑战[9].因此需在传统强化学习的基础上进行扩展,使之能够适应于多智能体环境.幸运的是,至少在学习期间,Agent 之间可以通过交互共享学习成果,改善学习性能[9].结合机器人足球比赛的特点,本文使用一种投票的多智能体强化学习方法来提高球队的协作能力,得益于强化学习的在线学习特点,球队能通过比赛不断提高自身能力.

3 一种基于投票的多智能体强化学习在足球机器人比赛中的应用

首先扩展 MDP 为多智能体 Markov 决策过程,定义为元组:(Ag, S, A, R, T),其中 $Ag, |Ag| = n$,表示参与协作的 Agent 集合,这里即为足球机器人的集合;$A = \times_{i \in Ag} A_i$ 是 Agent 为完成协作能采取的联合行为的集合;状态转移函数 T 表示在状态 s 采取联合行为 a 转移到状态 s' 的概率;R 为奖赏函数,对参与协作的所有 Agent 都相同,因为一次协作是他们共同确定的,对结果也应该共同分担.

本方法的主要思想是通过定义联合行为的集合然后通过投票综合考虑各球员的选择最终得出一种合适的供全体协作 Agent 采用的联合行为.

这里需要给出两个假定:一是参与协作的各 Agent 共享同一目标;二是 Agent 之间具有通信能力.本方法的一个关键概念是策略[10],定义:$\sigma = U_{k=1}^{n} \sigma_k$ 为 Agent 能够选择的策略集合;其中一个策略定义为:$\sigma_k = \{a_i^k | i = 1 \cdots n\}$ 也即一个联合行为,其中 a_i^k 为第 i 个 Agent 在策略 k 下采取的行为;可以看出,联合行为由一系列基本行为构成,表示在策略 k 下各球员采取的基本行为[11].其中基本行为可以是低层的动作选择,如一个队员选择传球,另一个队员选择移动到定点接球;也可以是抽象的角色选择,如一个队员主攻,一个队员助攻等.在机器人足球比赛中,不管是协作进攻,协作防守还是协作拦截,队友的正确跑位是完成一次漂亮协作的前提,本文的策略也以此为依据进行设计.

策略的引入带来两个好处,一是能以一种直接的方式综合各球员的选择并且可以只把注意力集中在那些高质量的策略上;二是使得状态-动作对空间大大减小,因为各球员都共享相同的高质量的联合行为,而这显然要比所有的状态-动作空间小很多.需要注意的是策略集合是预先定义好的,策略的好坏依赖于设计者.策略的质量影响学习过程进而影响系统整体性能.一种解决方法是定义大量的策略让球员在学习过程中选择逐步保留好的策略,去掉不好的[12];另一种方法是给出一定的初始策略,让球员在学习过程中逐渐进化,最终得到好的策略[13].

为了得到一个供所有协作成员执行的策略,需要采取某种方式综合考虑各个球员的选择.本文采用投票的方式.投票过程如下:每个球员根据自己的 Q 值表为它认为是最合适的策略投票,并且公布其投票,然后投票被综合,获得最高投票的策略将被确定为参与本次协作的所有球员下一步行为必须遵循的策略.数学

表达式如下：

$$\sigma_w = \arg\max y_{\sigma k} \tag{3}$$

其中，σ_w 为获得最高投票的策略，$y_{\sigma k}$ 为策略 k 获得的投票，计算方法为：

$$y_{\sigma k} = \sum_{i=1}^{n} d_{i\sigma k} \tag{4}$$

其中，$d_{i\sigma k}$ 为第 i 个球员为策略 σ_k 的投票.

之所以在强化学习中引入投票，主要考虑在足球机器人比赛中每个队员获得的环境信息不完整，对怎么完成共同的目标也不具备完备的知识[14,15]. 因此，把各个队员的知识综合考虑能获得更好的效果.

在本方法中，每个球员拥有自己独立的 Q 值表，并且将环境状态分为需要协作的状态和独立行为的状态. 当处于独立行为状态时，球员自主行为并且不修改 Q 值表；当进入一个需要协作的状态时，参与协作的球员根据自己的 Q 值表选择一个策略，然后通过投票的方式确定最终被每个球员都采用的策略. 奖赏函数只定义在需要协作的状态上，并且对于每个球员都一样，这进一步减小了状态空间. 不参与协作的球员不会收到任何奖励，也不会修改其 Q 值表.

基于投票的多智能体 Q 学习算法如下：

算法要求：一个策略集合，一个初始状态 s_0

(1) $s \leftarrow s_0$

(2) 循环执行

(3) 如果 s 是需要协作的状态，则：

(a) 使用 ε-贪心策略选择一个策略

(b) 公布自己的选择

(c) 接收其他 Agent 选择的策略

(d) 遵循投票方式计算得到要执行的策略

(e) 执行策略

(f) 获取奖励

(g) 转换到新状态 S'

(h) 更新 Q 值表

(4) 否则自主行为

这里有两个问题很重要，一是什么时候球员被认为是处于协作状态；二是投票过程按什么方式完成. 首先，协作必须至少有两个球员参与，对于第一个问题，当球员之间距离很近或者是一个球员距离其他已经处于协作状态的球员很近时，该球员可被认为处于协作状态. 在机器人足球比赛中，当控球队员遭遇拦截时，需要与距离最近的队友进行协作；或者当进入射门范围时，需要与其他队员协作以选择最佳射门位置. 对于第二个问题，有两种可选方式，一是设定一个 Agent 作为领导者，他收集其他 Agent 的投票计算出得票最高的决策并告知其他 Agent 执行；二是每个 Agent 都将自己的投票告知其他 Agent，同时接收其他 Agent 的投票结果，然后在本地计算出得票最高的决策并执行[16].

4 仿真实验及其分析

本文将多智能体强化学习应用于策略选择，使 agent 能通过比赛学会在不同环境状态下采用最佳的策略完成协作进攻、协作防守、协作射门等. 在机器人足球比赛中最重要的是赢球，但一般在比赛中进球是很少的，如果仅以此作为强化信号的惟一因素，学习的效果肯定很差，本文综合考虑战术是否完成、是否进球等因素确定强化值，具体可根据经验来确定.

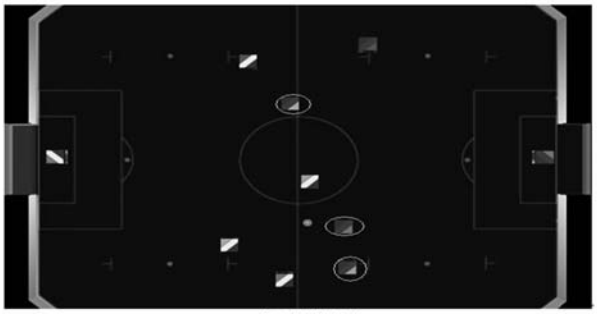
由于机器人足球比赛在动态复杂的环境中进行，对环境状态的表示至关重要，为减小状态空间，提高学习效率，本文使用区域代替坐标表示位置，将球场划分为若干个带编号的区域，区域的大小可以按经验或采用进化方法确定[17,18]. 这样，环境状态可由球所在的区域，球的速度，各区域的状态来确定，区域忽略球员身份，比如在某时刻对方球员 1 号，2 号，我方 5 号处于某区域内，另一时刻对方 3 号，4 号，我方 1 号在同一个区域内，那么这个区域在这两种状态下不作区分，这种对位置的离散化方式能极大地简化对环境的处理，这对机器人足球系统尤为必要. 前面提到，在学习过程中，队员之间的协作是动态组织的，需要时临时组织，完成后解散，本文采用一种默认和主动请求相结合的方式确定协作成员和协作时间，具体如下，当距球最近的球员与球的距离 d 在某范围内时该球员可以发起一次协作，以球为圆心，$2d$ 为半径，所得的圆经过和包含的区域内的所有队员默认参与协作，如果有队员根据自己的观察有自己的打算，可以请求退出，同时范围之外位置更好的队友也可以主动请求参与协作，这种方式极大地提高了灵活性的同时又兼顾到队员自身拥有的局部知识，学到的能力将能更好地应用于环境.

对不同的联合行为结果应给予不同的奖励值，以之前统计的球进禁区次数和射门次数之比（约 2.612，为简单起见取 2.5）的倒数作为本次的奖罚比，故本文实验按如下方式处理，如果在当前状态下本次协作完成了一次射门，不论是否进球，因为有可能对方守门员挡住了，都给予所有参与协作的队友 + 100 的奖励值；如果失败，如丢球，则给予 – 40 的奖励值以惩罚其错误的决策，考虑机器人本身硬件限制和足球比赛的复杂性，大多数协作可能是部分成功的情况，本实验中对此只给予很小的惩罚值 0.002，同时折扣因子 γ 取 0.9. 机器人足球 SimuroSot 5vs5 仿真比赛的结果证明这是合理的.

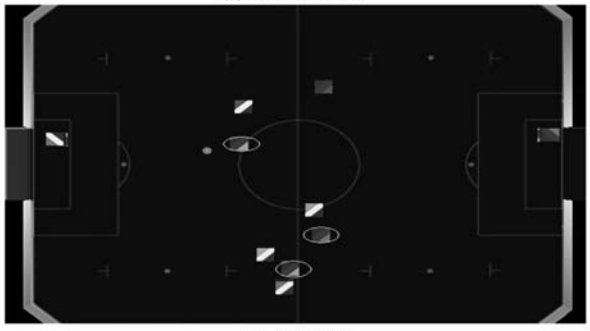
图 1 的 a，b 显示了蓝队三个用白圈标注的队员使用本文方法学到的战术完成的一次进攻.

在训练过程中，本文选择 2008 年本校的冠亚军和

2007 年全国比赛四强之一的程序分别作为对手,三个程序分别采用的模糊基本动作集和遗传策略,但均为集中式控制机器人组.前期球队的成绩没有什么大的进步,主要原因是系统的状态空间较大,学习初期 Q 值的影响不准确,而且比赛的随机性较大,对比赛结果的影响要大于学习的效果.前面提到,本文方法会使状态空间随着训练逐步减小,策略也会逐步优化,大约经过 80 场比赛以后,系统性能开始逐步提高;在约 150 场比赛后,平均进球数能够稳定在 40 个左右,原因是球队已经习惯了对手,动作策略维持在一个相对稳定状态,结果如图 2 所示。



(*a*) 进入协作状态



(*b*) 完成协作

图1　三个队员的一次协作进攻

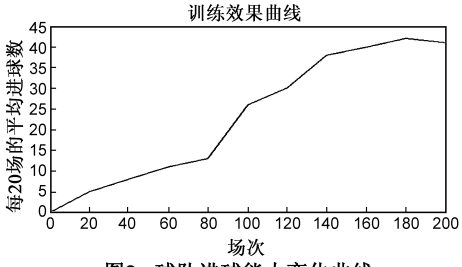


图2　球队进球能力变化曲线

同时,本文统计了分别采用集中式决策方法和本文方法的两支球队在仿真比赛中成功完成的协作场景所花费的平均时间,结果如图 3 所示.

图中横坐标为场景次数,纵坐标为完成协作所花费的平均时间,蓝色实线表示采用本文方法的曲线,绿色虚线表示采用集中式决策方法的实验曲线.足球机器人仿真比赛对实时性要求很高,不管是协作进攻还是拦截,都应在以秒为数量级的时间内完成,否则在瞬息万变的赛场形势中策略将变得毫无意义.由上图比较可以发现,本文方法在时间上比经典的集中式控制方法在实时性上具有明显的优点.

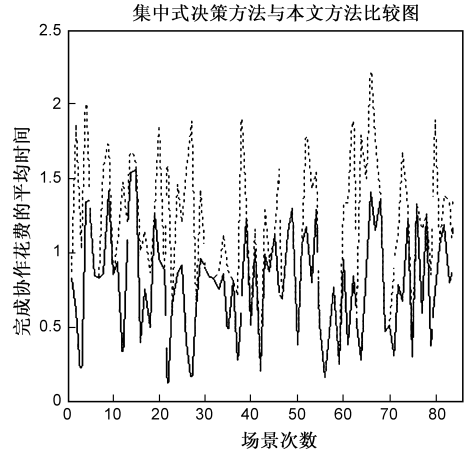


图3　完成协作所花费的平均时间对比曲线

5　结论

本文针对机器人足球系统提出了一种多 Agent 强化学习方法以解决在动态环境中多个机器人获取联合行为的问题.增强系统的适应性是多 Agent 强化学习的根本目的.但强化学习的实现必定要受到相关学习技术和领域问题的限制,例如 Q 学习对于比较简单的问题易于实现,但是如果应用于大状态行为空间,则有可能出现所谓的"维数灾难",因为整个状态行为空间将随着 Agent 个数的增加呈指数级增长.本文通过对球场的划分,动态地组织协作有效解决了行为空间问题;通过定义策略表示联合行为,从而将协作问题转化为对策略的学习;采用投票的方式综合考虑各 Agent 的决策,遵循票数最高的策略达到协作完成任务的目的.由实验可以看出,系统在实时性和进球能力上有了明显提升,实验的奖励值采用的是之前积累的比赛统计数据,在以后的研究中可以考虑采用收敛性更好的方法.

参考文献:

[1] Milind Tambe, Jafar Adibi, Yaser Al-Onaizan, Ali Erdem Gal A Kaminka, Stacy C Marsella, Ion Muslea. Building agent teams using an explicit teamwork model and learning[J]. Artificial Intelligence, 1999, 110(4): 215 – 239.

[2] M Riedmiller, T Gabel, J Knabe, H Strasdat. Brainstormers 2D-Team Description 2005[DB/OL]. http://panmental.de/papers/BSTeam05.pdf, 2005.

[3] R S Sutton, A G Barto. Reinforcement Learning: An Introduction[M]. Massachusetts: MIT Press, 1999. 39 – 46.

[4] SINGH S. Agents and Reinforcement Learning[M]. San Mateo, CA: Miller Freeman publish Inc, 1997. 42 – 77

[5] C Watkins, P Dayan. Q-learning[J]. Machine Learning, 1992, 324(8): 279 – 292.

[6] C Claus, C Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems[A]. Proc of the 15th National Conference on Artificial Intelligence[C]. Menlo Park CA: American Association for Artificial Intelligence, 1998. 746 – 752.

[7] Zhao Jian-ming, Mao Xin-jun, Wang Ji. Developing multi-agent systems with dynamic binding mechanism[A]. IAT′06 Proceedings of the IEEE/WIC/ACM international conference on Intelligent Agent Technology[C]. Washington DC USA: IEEE Computer Society, 2006. 12 – 24.

[8] Sarit Kraus, et al. Multiagent negotiation under time constrain [J]. Artificial Intelligence, 1995, 75(6): 297 – 345.

[9] Myoung Hwan Choi, Bum Hee Lee. A real time optimal load distribution for multiple cooperating robots[A]. Proceedings of 1995 IEEE International Conference on Robotics and Automation[C]. Nagoya, 1995. 1211 – 1216.

[10] David V Pynadath, Milind Tambe. The communicative multiagent team decision problem: Analyzing teamwork theories and models[J]. Journal of Artificial Intelligence Research. 2002 (16): 389 – 423.

[11] Caloud, P Wonyun Choi, Latombe. J-C Le Pape. C Yim M. Indoor automation with many mobile robots[A]. Proceedings. IRO ′90. IEEE International Workshop on Intelligent Robots and Systems ′90.′ Towards a New Frontier of Applications [C]. Ibaraki, 1990. 67 – 72.

[12] Franko Sore, Hugh Rudnick, et al. Definition of an efficient transmission system using cooperative games theory[J]. IEEE Transactions on Power Systems, 2006, 21(4). 1484 – 1492.

[13] Rohit Gupta, Arun K Somani. Game theory as a tool to strategize as well as predict nodes behavior in peer-to-peer networks [A]. Proceedings of 11th International Conference on Parallel and Distributed Systems[C]. Washing D C IEEE Computer Society, 2005. 244 – 249

[14] LI Ning, GAO Yang, LU Xin, CHEN Shi-Fu. A learning agent based on reinforcement learning[J]. Computer Research and Development, 2001, 38(9): 1051 – 1056.

[15] CAI Qing-sheng, ZHANG Bo. An agent team based reinforcement learning model and its application[J]. Journal of Computer Research and Development, 2000, 37(9): 1087 – 1093.

[16] Kenneth S Rubin, Adele Golberg. Object behavior analysis [J]. Communications of the ACM, 1992, (9): 48 – 62.

[17] C Guestrin. Planning Under Uncertainty in Complex Structured Environments[D]. USA: Department of Computer Science of Stanford University, 2003.

[18] Martin L Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming[M]. United States: Wiley-Interscience, 2005. 302 – 368

[19] 周兰凤,洪炳熔.用基于知识的遗传算法实现移动机器人路径规划[J].电子学报,2006,34(5):911 – 914.
ZHOU Lan-feng, HONG Bing-rong. A knowledge based genetic algorithm for path planning of a mobile robot[J]. Acta Electronica Sinica, 2006, 34(5): 911 – 914.

作者简介:

刘春阳 男,1978 年 10 月出生于河北唐山.现为华北电力大学控制与计算机工程学院教师,北京科技大学信息工程学院博士研究生,主要研究方向包括:智能理论、智能机器人等.
E-mail: liuchuny@ncepu.edu.cn

柳长安 男,1971 年 12 月出生于黑龙江拜泉.现为华北电力大学控制与计算机工程教师,教授,华北电力大学智能机器人研究所所长,主要研究方向包括:智能控制,机器人技术,嵌入式技术等. E-mail: liuchangan@ncepu.edu.cn