Robocup半场防守中的一种强化学习算法

冯 林[1], 李 琛[1,2], 孙 焘[1]
(1.大连理工大学, 辽宁 大连 116024;
2.海军91439部队,辽宁 大连 116041)

摘 要:Robocup仿真比赛是研究多Agent之间协作和对抗理论的优秀平台,提高Agent的防守能力是一个具有挑战性的问题。为制定合理的防守策略,将Robocup比赛中的一个子任务——半场防守任务分解为多个一对一防守任务,采用了基于Markov对策的强化学习方法解决这种零和交互问题,给出了具体的学习算法。将该算法应用到3D仿真球队——大连理工大学梦之翼(Fantasia)球队,在实际比赛过程中取得了良好效果。验证了采用Markov零和对策的强化学习算法在一对一防守中优于手工代码的结论。

关键词:Robocup; 强化学习;Markov对策;零和对策
中图分类号:TP242.6 **文献标识码**:A **文章编号**:1673-629X(2008)01-0059-04

A Reinforcement Learning Method for Robocup Soccer Half Field Defense

FENG Lin[1], LI Chen[1,2], SUN Tao[1]
(1.Dalian University of Technology,Dalian 116024, China;
2.Navy Forces 91439,Dalian 116041,China)

Abstract:Robocup soccer simulation is an excellent platform in which collaboration and counterwork among multi-agent are studied. It is a challenging problem to improve agent's defense ability. In order to design reasonable defending policy, decompose a subtask, half field defense, into some one-vs-one defense subtask and pose it as a problem of zero-sum Markov games. In this paper, a reinforcement learning method based on Markov game is developed and implemented in 3D simulation soccer team—DUT Fantasia. In real matches, this arithmetic is approved to be efficient and better than manual-coding in one-vs-one defense subtask.

Key words:Robocup; reinforcement learning; Markov game; zero-sum game

0 引 言

未来50年,人工智能领域的主要问题是"多主体动态不可预测环境中的问题求解",作为其基本问题,训练机器人进行足球比赛正成为研究热点之一,机器人足球世界杯(Robocup)[1]为此提供了一个平台。

Robocup仿真比赛不考虑硬件条件的限制,用软件平台尽量模拟真实世界中的足球比赛情况,强调以有限的能力(认知能力有限,行为能力有限,通讯能力有限)和条件在实时、动态的环境中活动,感知环境,相互合作,并在对抗性的竞赛中获得胜利。因此,Robocup仿真比赛被认为是研究多Agent之间协作和对抗理论的优秀的平台。

收稿日期:2007-07-04
基金项目:国家自然科学基金(50575031)
作者简介:冯 林(1969-),男,博士,教授,研究方向为图像压缩、配准及融合、演化算法。

国内外的研究人员针对Robocup仿真比赛中的各种技术应用作了很多工作,Peter Stone将分层学习应用到Robocup仿真球队CMUnited的架构中[2],清华大学用Q学习训练Agent的踢球和带球等个人技术[3],德国卡尔斯鲁厄大学的Robocup仿真球队Brain-Stormer则利用神经网络训练球队的技战术[4]。这些应用在实际的Robocup比赛中都取得了很好的效果,但是上述工作大多集中于Agent个体能力的学习或者多Agent之间的协作关系,较少涉及Agent之间的对抗问题。

文中将Robocup仿真比赛中的一个子任务——半场防守作为研究重点,针对一对一防守这一对抗性问题,应用基于Markov对策的强化学习方法,制定防守策略,并尝试将其应用在3D仿真球队大连理工梦之翼(Fantasia)中。

在所进行的训练和比赛实验中,该防守策略取得了较好的效果。

1 强化学习

强化学习解决这样的问题:一个能够感知环境的自治 Agent,怎样通过学习选择能达到其目标的最优动作。Agent 与环境的交互接口包括行动(Action)、回报(Reward)和状态(State),学习过程如图 1 所示。

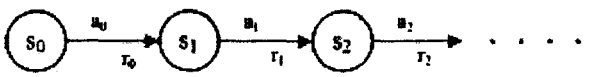

图 1 强化学习的学习过程

Agent 为了完成任务,必须知道采取某个策略而导致出现的某个状态对该 Agent 所产生的长期回报,而不是其立即回报。长期回报必须经过一定时间的延迟才能获得,Agent 在 t 时刻获得的长期回报可以用式(1)来表示:

$$V^{\pi}(s_t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots = \sum_{i=0}^{\infty} \gamma^i r_{t+i} \tag{1}$$

其中 $V^{\pi}(s_t)$ 是 s_t 状态下采取策略 π 而产生的长期回报,γ 是延迟折算因子。

如果 t 时刻之后的环境状态只取决于 t 时刻的状态 s_t 和 t 时刻之后的动作,而与 t 时刻之前的状态和动作无关,则该学习过程可看作是一个 Markov 决策过程(MDP),那么可以用 Q 学习的方法来学习最优策略,其学习方法如下所示:

$$V(s) = \max_{a \in A} Q(s,a) \tag{2}$$

$$Q(s,a) = R(s,a) + \gamma \sum_{s' \in S} T(s,a,s') V(s') \tag{3}$$

其中 A 是进行学习的 Agent A 的动作集,S 是所有可能的状态集,T 是状态转移函数,s' 是当前状态 s 下采取动作 a 所得到的后继状态,$Q(s,a)$ 是当前状态 s 下采取最优动作 a,以后一直遵循最优策略 π^* 时所获得的折算长期回报。Q 学习的目标就是通过不断学习而选择一个最优策略 π^*,使得总的折算长期回报最大[5]。

2 Markov 对策

Markov 对策可以用以下多元组表示:状态集 S,Agent 的个数为 n,$A_1,\cdots,A_n$ 分别是每个 Agent 的动作集,状态转移函数 $T: S \times A_1 \times \cdots \times A_n \to \mathrm{PD}(S)$,其中 $\mathrm{PD}(S)$ 为状态分布,联合回报函数 $R_i: S \times A_1 \times \cdots \times A_n \to \Re$,其中 R_i 是 Agent i 在状态 s 下,所有 Agent 采取动作后获得的回报[6]。

如果环境中只存在两个互相竞争的 Agent A 与 O,他们的动作集分别为 A 和 O,则 A 所获得的折算长期回报 $Q(s,a)$ 可以扩展为 $Q(s,a,o)$,即当前状态 s 下,A 采取动作 a 而 O 采取动作 o 时所获得的长期回报。这种简单的一对一竞争交互,可看作是二人零和对策问题,则 Q 学习的基本方程可扩展为式(4)和(5):

$$V(s) = \max_{\pi \in PD(A)} \min_{o \in O} \sum_{a \in A} Q(s,a,o)\pi_a \tag{4}$$

$$Q(s,a,o) = R(s,a,o) + \gamma \sum_{s' \in S} T(s,a,o,s') V(s') \tag{5}$$

式(4)中,对手 Agent O 选择对 A 最不利的动作 o,π_a 为动作集 A 上的分布,该式的含义是在对手 Agent 选择对自己最不利的动作的情况下,自己选择最有利的动作所获的的回报,这种方法叫做"极大极小 Q 学习"法[7]。

3 Robocup 半场防守中的学习模型

3.1 Robocup 中的半场防守问题

虽然 Robocup 仿真比赛可以作为研究多 Agent 以及机器学习的优秀平台,但是其实时性和动态性以及状态空间巨大等特点,导致无法将强化学习等方法直接应用来解决所面临的问题,因此需要对问题进行分解,将待解决的问题逐步细分成若干相对比较简单的子问题。一场 Robocup 仿真比赛的过程是 22 个 Agent 在仿真的足球比赛环境中进行协作和对抗的过程,文中将这一过程按照一场真正的足球比赛所要解决的问题分解为如图 2 所示的一些子问题。

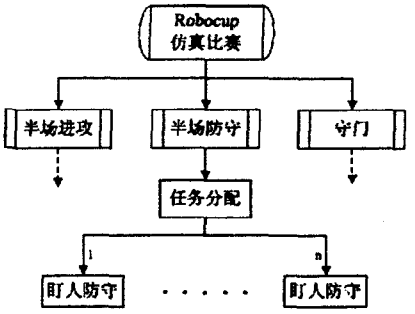

图 2 Robocup 中的问题分解

如果半场防守采用人盯人防守策略,则其可进一步分解为防守任务分配以及一对一盯人防守,防守任务的分配可以采用各种 Agent 协作理论来解决,一对一防守则可以采用基于 Markov 零和对策的强化学习方法来解决,为了简单明了地验证该学习方法在半场防守中的实际应用效果,采用贪心算法实现任务分配,而学习算法只在解决一对一防守问题时应用。

3.2 防守任务分配

在 Robocup 比赛中,球进入防守区域后,如果防守一方无法控球,半场防守策略即被触发。此时应该首先根据场上形势确定需要参与攻防对抗的每一方的 Agent 个数 n(由于采用一对一盯人防守策略,参与攻防对抗的防守 Agent 与进攻 Agent 的个数相同,都是 n

),然后用贪心算法将防守 Agent 与进攻 Agent 匹配,匹配后的攻防场景如图 3 所示。

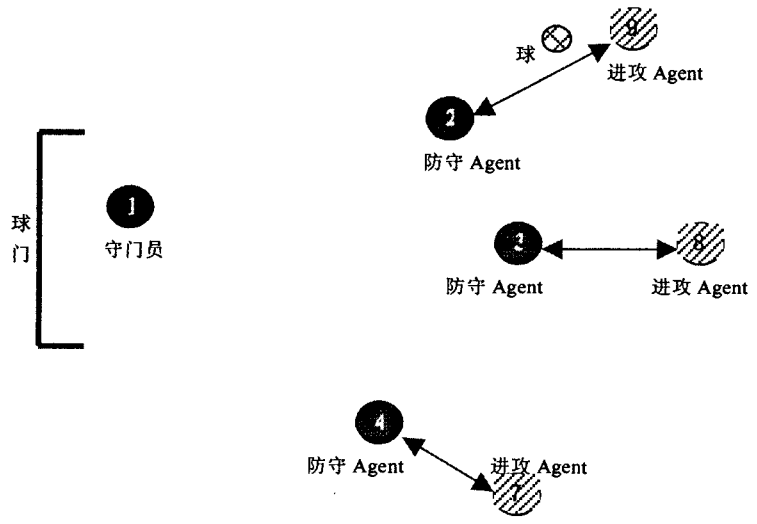

图 3　半场攻防对抗场景

在进行任务分配时,根据进攻 Agent 与球门的距离和角度确定其危险度,根据防守 Agent 与进攻 Agent 的距离和角度选择防守 Agent 与进攻 Agent 进行匹配。在该场景中,处于防守区域内参与攻防对抗的 Agent 共有 6 个(守门员不参与盯人防守),形成了 3 个一对一攻防对抗任务(用箭头连接的进攻和防守 Agent)。

3.3　一对一盯人防守

3.3.1　环境建模

针对如图 4 所示的一对一防守这一特定场景,Agent 所处的环境可描述如下:

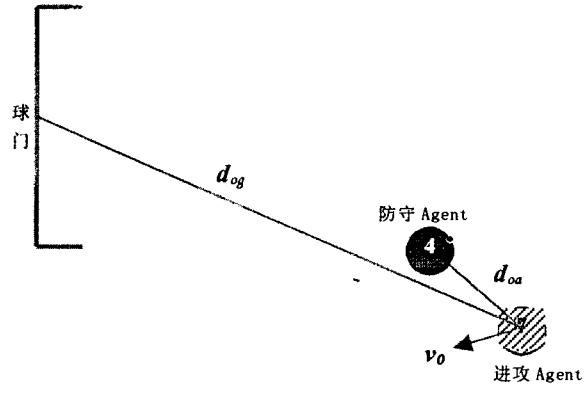

图 4　一对一盯人防守场景

d_{og}:进攻 Agent 到球门中点的距离矢量,包括大小和方向。

d_{oa}:进攻 Agent 与防守 Agent 的距离矢量,包括大小和方向。

v_o:进攻 Agent 的速度矢量,包括大小和方向。

在任意时刻,防守 Agent 关于当前状态的立即回报遵循下式:

$$r = f(\boldsymbol{d}_{og}, \boldsymbol{d}_{oa}, \boldsymbol{v}_o) \tag{6}$$

防守 Agent 和进攻 Agent 的动作集均为 Move(dir, power),其中 dir 的取值范围是(0 ~ 360°),power 的取值范围是(0 ~ MaxPower),MaxPower 和 $T(s,a,o,s')$ 均可由 Robocup 仿真比赛平台获得。在具体学习过程中,可根据计算精度和效率选择一定步长取 dir 和 power 的一些离散值点,构成 Agent 的动作集 A 和 O。

3.3.2　学习算法

根据以上的环境建模,学习的基本算法如下:

(1)初始化参数,对所有 $s \in S, a \in A, o \in O, Q(s,a,o) = 1, V(s) = 1, \alpha_0 = 1.0, \pi(s,a) = 1/|A(s)|$;

(2)选择一些特定场景,初始化实际环境;

(3)选择动作,在当前状态 s 下,以概率 $\pi(s,a)$ 选择动作集 A 中的 a 执行;

(4)学习最优策略 π^*。

计算立即回报:观察对手,自己以及球的状态,根据式(6)计算 r。

修正学习率:$\alpha = \alpha_0/n(s,a,o)$,其中 $n(s,a,o)$ 为行为状态对 (s,a,o) 出现的次数。

更新 Q 值:$Q(s,a,o) = (1-\alpha)Q(s,a,o) + \alpha(r + \gamma V(s'))$。

选择最优策略 $\pi^*(s,\cdots)$:使得 $\min_{o' \in O} \sum_{a' \in A} \pi'(s,a')Q(s,a',o')$ 取得最大值,令 $V(s)$ 等于次最大值。

(5)如果当前状态是终止状态,则转向步骤(2),初始化场景;否则转向步骤(3),继续选择动作。

4　实验结果

将 3.3.2 的一对一学习算法应用在 Robocup 仿真比赛的 3D 平台上,让一个进行学习的防守 Agent 与 2005 年获得全国比赛季军的大连理工梦之翼队的前锋进行一对一的攻防训练。场景初始化为防守 Agent 位于禁区线上的随机位置,进攻 Agent 位于中线上的随机位置,进攻 Agent 的目标为带球向禁区推进,当出现如下状态时,该场景结束:

(1)进攻 Agent 成功带球推进到禁区,防守 Agent 获得回报 -1;

(2)球出界,防守 Agent 获得回报 0;

(3)防守 Agent 断球,防守 Agent 获得回报 1;

(4)迭代次数到达 300 个仿真周期。

回报函数 r 的值满足：$|f(d_{og}, d_{oa}, v_o)| < 0.005$，动作集 Move(dir, power)离散为 24 * 5 个值。如果在 500 个仿真周期内，进攻 Agent 没有成功带球推进到禁区或者防守 Agent 成功断球，则认为防守成功，否则防守失败。经过 30000 个场景的训练，得到了基本稳定的防守策略。

将学习后的防守策略应用到 3D 仿真球队 Fantasia 中，并与 2005 年全国比赛的冠军浙大求是队、亚军科大蓝鹰队以及 2005 年世界杯冠军 Aria 分别进行了 100 场比赛(每场比赛 6000 个仿真周期)，每场比赛对手带球推进到禁区的次数和失球数如表 1 和 2 所示。

表 1 对手带球推进到禁区的平均次数

	随机策略	手工代码	学习代码
浙大求是队	16.62	12.37	4.18
科大蓝鹰队	14.35	9.28	2.49
Aria	18.21	12.56	4.75

表 2 平均失球数

	随机策略	手工代码	学习代码
浙大求是队	3.13	1.21	0.33
科大蓝鹰队	2.51	0.82	0.15
Aria	3.86	1.24	0.38

应用该防守策略的 3D 仿真球队 Fantasia 参加了 2006 年 6 月份在德国不来梅举行的 Robocup 世界杯和 2006 年 10 月份在苏州举行的 Robocup 全国比赛暨中国公开赛 3D 仿真组的比赛。在不来梅进行的 10 多场世界杯比赛中，Fantasia 只丢了一个球，输了一场比赛，并最终获得世界杯第 11 名；在苏州参加的近 20 场比赛中，Fantasia 一球未失，最终以不败战绩囊括全国比赛和中国公开赛的冠军。

5 结论与展望

针对 Robocup 半场防守这一对抗性问题，将多对多的半场防守任务分解为多个一对一防守的子任务，并用 Markov 零和对策模型描述该任务，研究了基于 Markov 对策的强化学习方法在一对一防守中的应用。实验和比赛结果表明，该方法取得了优于手工代码的应用效果。由于篇幅有限，在防守任务分配问题上未作详细讨论，实际上该环节还有改进的余地，例如可以采用“匈牙利算法”[8]或者基于阵型的角色分配[9]等解决进攻 Agent 和防守 Agent 的一对一匹配问题。

参考文献：

[1] Kitano H, Tambe M, Stone P, et al. The RoboCup Synthetic agent challenge97[C]// In Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence. Nagoya: [s. n.], 1997: 24 - 29.

[2] Stone P. Layered Learning in Multi - Agent Systems[D]. Pittsburgh, PA, USA: Computer Science Department, Carnegie Mellon University, 1998.

[3] Yao Jinyi, Chen Jiang, Cai Yunpeng, et al. Architecture of Tsinghua Aeolus[C]// In: Birk A, Coradeschi S, Tadokoro S eds. Robocup 2001: Robot Soccer World Cup V. Heidelberg: Springer - Verlag, 2002.

[4] Riedmiller M, Braun H. A direct adaptive method for faster back - propagation learning: The RPROP algorithm[C]// In Ruspini H ed. Proceedings of the IEEE International Conference on Neural Networks (ICNN). San Francisco: [s. n.], 1993: 586 - 591.

[5] 张汝波. 强化学习理论及应用[M]. 哈尔滨：哈尔滨工程大学出版社，2001.

[6] Owen G. Game Theory[M]. 2nd Edition. Orlando, FL, USA: Academic Press, 1982.

[7] Littman M L. Markov games as a framework for multi - agent reinforcement learning [C] // In Proceedings of the Eleventh International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann, 1994: 157 - 163.

[8] 常庭懋，韩中庚. 用“匈牙利算法”求解一类最优化问题[J]. 信息工程大学学报，2004，5(1)：60 - 62.

[9] Stone P, Veloso M. Task decomposition. dynamic role assignment, and low - bandwidth communication for real - time strategic teamwork[J]. Artificial Intelligence, 1999, 110(2): 241 - 273.

(上接第 58 页)

Group, 1984.

[2] Yin De - Sheng, Wang Guo - Yin, Wu Yu. A Self - learning Algorithm for Decision Tree Pre - Pruning[C]// Proceedings of the Third International Conference on Machine Learning and Cybernetics. Shanghai: [s. n.], 2004.

[3] 苗夺谦，王 珏. 基于粗糙集的多变量决策树的构造方法[J]. 软件学报，1997，8(6)：425 - 431.

[4] 史忠植. 知识发现[M]. 北京：清华大学出版社，2002.

[5] 胡学钢，张冬艳. 一种新的基于粗集的决策树构造算法[J]. 计算机科学，2005，32(8A)：7 - 9.

[6] Pawlak Z. Rough Set Approach to Multi - Attribute Decision Analysis[J]. European Journal of Operational Research, 1994, 72(3): 443 - 459.

[7] 曾黄麟. 粗集理论及其应用[M]. 重庆：重庆大学出版社，1998.

[8] 沙慧新. 基于知识粗糙度和拓展属性约简的若干智能挖掘算法的研究[D]. 福州：福州大学，2004：13 - 17.