Scientific Journal of Control Engineering
Volume 6, 2016, PP.42-46

Research on the RoboCup defensive strategy

Junpeng ZHANG[#1], Wei CHEN[1], Haohui HUANG[1]

1.College of Automation, Guangdong University of Technology, Guangzhou, Guangdong, 510006, China

[#]Email: 2682390094@qq.com

Abstract

Robocup soccer simulation is an excellent platform in which collaboration and counterwork among multi-agent are studied. The height of the team is decided by the Robocup defensive. How to improve the Robocup defense is very challenging problem. I take the game in order to better understand and training Agent is decomposed into multiple tasks one-on-one. Based on improving the Q algorithm I am learning to solve the defensive problems in the game. Applying the algorithm to the team of Guangdong university of technology, good results have been achieved in the actual game, Based on the improved Q learning is verified on the defensive works as an ideal.

Keywords:Robocup;Reinforcement Learning; Zero-sum Game

RoboCup 防守策略研究

张俊朋[1]，陈玮[2]，黄浩辉[3]

1. 广东工业大学 自动化学院，广东 广州 510006

摘　要：Robocup 仿真赛是研究 Agents 之间的对抗和协作理论比较理想的平台。球队的高度是由 Robocup 的防守来决定，如何提高 Robocup 的防守是一个很有挑战的问题。为了更好的理解和训练我把比赛中的 Agent 分解成多个一对一的任务。采用基于改进行的 Q 算法来解决比赛中的防守问题。将该算法应用到广东工业大学的球队上，在实际比赛中取得了很好的效果。验证了基于改进的 Q 算法在防守起到比较理想的作用。

关键词：Robocup；强化学习；零和对策

引言

未来的几十年里，人工智能领域，智能体与多智能体系统（Multi-Agent System）成为研究热点，这些都是围绕着多主体在未知环境中做出自我决策和自我修正的问题[1]因此训练机器人进行足球比赛成为当前研究的热点之一，机器人世界杯（Robocup）[1]为我们提供了一个很好的研究平台。

RoboCup 为智能行为决策和强化学习提供了一个标准的研究平台，其核心问题是一个巨大的连续状态、动作及观察空间的多智能体学习和实时决策问题[2]。自 1997 年 RoboCup 比赛首次举办，16 年来它吸引和激励了来自世界各地的很多研究者，关于 RoboCup 的文章已有数百篇。在最早的比赛中，机器人的策略只是手工的分支判断与选择，没有学习和智能可言。16 年过去了，球队的智能化水平已大大提高，但是当初它提出的问题仍然代表着当今人工智能和机器人学的前沿，多智能体环境下的智能决策规划仍是研究的重点和难点[3]。国内外研究 Robocup 仿真比赛的很多，但是大部分都是强调的 Agent 本生学习能力和多 Agent 之间的协调能力，很少提及 Agent 之间的对抗问题[4]。

1　Q 算法与其改进

2.1　算法

Q 学习算法由 C.Watkins 于 1989 年在其论文“Learning from delayedrewards [5]”中首次提出，该算法是动

态规划的有关理论及动物学习心理学的有力相互结合，以求解具有延迟回报的序贯优化决策问题为目标。在 Q 学习算法中根据 TD 算法对 Markov 决策过程的行为值函数进行迭代计算，其迭代计算公式为:

$$Q(s_t,a_t)=(1-l)Q(s_t,a_t)+l[r(s_t,a_t)+\gamma \max_{a_{t+1}} Q(s_{t+1},a_{t+1})] \quad (1)$$

其中 $Q(s_t,a_t)$ 为 Markov 决策过程在时刻 t 的状态一行为对，S_{t+1} 为 t+l 时刻的状态，$r(s_t,a_t)$ 为 t 的回报，l>0 为学习因子。同时还需要给定有限离散状态和行为空间 Markov 决策过程的状态集 S 和行为集 A，扣总回报目标函数，其中折扣因子为 Y;以表格形式存储的行为值函数估计值 Q(s, a)及行为选择策略 π 。

1.初始化行为值函数估计和学习因子 l，初始化 Markov 决策过程的状态，令时刻 t=0;

2.循环，直到满足停止条件为止;

a)对当前状态 s_t ，根据行为选择策略决定时刻 t 的行为 a_t ，并观测下一时刻的状态 s_{t+1}

b)根据迭代公式((1-1)更新当前状态一行为对的行为值函数的估计值:

c)跟新学习因子，令 t=t+1，返回 a。

2.1 算法的改进

Q 学习算法在强化学习领域受到普遍关注，因此，针对该算法的改进也层出不穷。其中 Peng 在 1996 年提出了增量式多步 Q 学习[6]，即 Q（ λ)算法，它结合了 Q 学习和 TD（ λ ）回报的思想，利用将来无限多步的信息更新当前 Q 函数。无论 Q 学习还是 Q（ λ)算法有存在着问题:Q 学习利用了一步信息，更新的速度慢，预见能力不强，而 Q（ λ)算法要对大量的状态一动作对，当状态一动作空间规模很大时计算量较大，学习效率不高。为了平衡这两个方面的问题，采用有限多步信息进行更新的思想，即多步 Q 学习算法。它利用 k 步的信息更新当前的 Q 值，具有多步预见能力，同时能降低计算复杂度[7]。这里的预见能力是指智能体更新当前状态一动作对的 Q 值时，考虑了将来若干个状态一动作对的影响，反映了智能体对未来的思考，从而使当前决策更加理性。k=1 时，它退化为 Q 学习算法，k=∞ 。时又演变为 Q（ λ)算法，因此它是这两种方法的结合[8]。同时在形式上又是两种方法的统一。

时刻的校正 n 步截然回报依然下式所示，考虑到对于 0< λ <1。

$$(1-\lambda)\sum_{j=1}^{k-1}\lambda^{j-1}+\lambda^{k-1}=1 \quad (2)$$

令 $\sum_{j=1}^{0}\lambda^{j-1}=0$ ，则（2）式当 k=1 时也成立。

定义 1 k 步的 TD(λ)回报定义为 r_t^n ,(n=1,2,3...k)的加权平均值，即

$$r_t^{\lambda,k}=(1-\lambda)\sum^{k-1}\lambda^{j-1}\bullet r_t^j+\lambda^{k-1}\cdot r_t^k \quad (3)$$

对于定义 1 中的 $r_t^{\lambda,k}$ ，存在

$$r_t^{\lambda,k}=r_t+\gamma\cdot v_{t-1}(x_{t+1})+\sum^{k-1}(\lambda r)^i\cdot[r_{t+i}+\gamma\cdot v_{t+i-1}(x_{t+i-1})-v_{t+i-2}(x_{t+i})] \quad (4)$$

可以有数学归纳发证明。如果学习效率 α 的值比较小，那么 Q 值的调整就比较慢，相应的 v 值调整也比较慢，由此可以得到

$$e_t=r_t+\gamma\cdot v_{t-1}(x_{t+1})-v_{t-1}(x_t) \quad (5)$$

$$e_t^{'}=r_t+\gamma\cdot v_{t-1}(x_{t+1})-Q_{t-1}(x_t,a_t) \quad (6)$$

由（5）和（6）可以得出

$$r_t^{\lambda,k}-Q_{t-1}(x_t,a_t)=e_t^{'}+\sum^{k-1}(\lambda r)^i\cdot e_{t+i} \quad (7)$$

$$\text{如果 } x\neq x_t \text{ 且 } a\neq a_t \quad Q_t(x_t,a_t)=\begin{cases} Q_{t-1}(x_t,a_t)+a\cdot[e_t^{'}+\sum_{i=1}^{k-1}(\lambda\gamma)^i\cdot e_{t+1}] \\ Q_{t-1}(x_t,a_t) \end{cases} \tag{8}$$

因此可以的出多部 Q 多步算法：

1.初始化 Q 数组

2.初始化状态 x 和数组 X，A，E, $E^{'}$，i=1

3.依据一定的策略选择状态 x 上的动作 x，执行动作 a，得到回报 r，状态转移到 x+1

4.根据（1-4）（1-5）计算的 e 和 $e^{'}$,并更新数组 X，A，E 和 $E^{'}$

5.若 i≤k,则 Q（x,a）=Q(x,a)+a·e':否则，Q（x[0],A[0]）=Q(X[0],A[0])+a· ∀

其中 $\forall = E^{'}[0]+\sum^{n-1}(\lambda\gamma)\ \cdot E[m]$

6.x←x+1，i←i+l，如果步数 i 达到设定值，算法结束;否则转到步骤 3

2 Robocup 防守模型

2.1 设计思想

Robocup 仿真比赛平台，就是一个多智能体的平台。每个 Agent 作为一个独立智能体，在不直到或者不完全知道比赛环境的前提下，通过对自身周围状态的信息进行详细的分析，做出相应反应。与此同时，与各个 Agent 间进行不同的判别：与队友之间的相互合作，与对手之间进行各种对抗[9]。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 1 所示的一些子问题。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[10]，一对一防守则可以采用基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[11]。

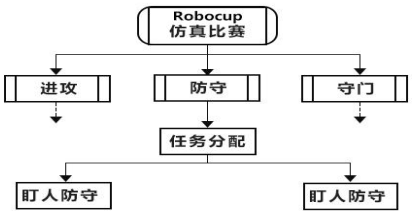

图 1

2.1 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[12]。(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）进攻和防守的图分配图 2 所示。

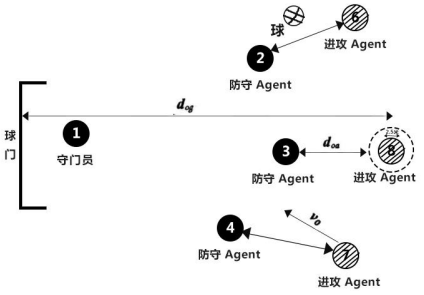

图 2

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 A-gent 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 A-gent)[13]。

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 $r = f(d_{og}, d_{oa}, v_0)$ 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 一 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[14]。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[15]。

2.1 算法实现

根据以上模型，实现算法:

1.初始化参数 $s \in S, a \in A, o \in O, Q(s,a,o)=1, V(s)=1, a_0=1.0, \pi(s,a)=1/|A(s)|$

2.选择一些特定场景，初始化实际环境。

3.依据一定的策略选择状态 S 上的动作 s，以概率 $\pi(s,a)$ 执行动作 a，得到回报 r，状态转移到 s+1

4.根据（1-4）（1-5）计算的 e 和 e',并更新数组 Q，S，E 和 E'

5.若 i≤k,则 Q（s,a）=Q(s,a)+a·e':否则，Q（s[0],A[0]）=Q(s[0],A[0])+a· ∀

其中 $\forall = E'[0] + \sum^{n-1} (\lambda\gamma) \cdot E[m]$

6 .s←s+1，i←i+l，如果步数 i 达到设定值，算法结束;否则转到步骤 3

3 实验结果

将 2.3 的一对一学习算法应用在 Robocup 仿真比赛的 2D 平台上，让一个进行学习的防守 Agent 与 2015 年获得全国比赛亚军的广东工业大学 GDUT_TiJi 的球员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。

回报值满足: $f|(d_{og}, d_{oa}, v_0)| < 0.48$, 动作集 Move (dir, power)离散为 30×5 个值。如果在 600 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 40000 个场景的训练，得到了基本稳定的防守带略[16]。

将学习后的防守策略应用到 2D 仿真球队 GDUT_TiJi 中，并与 2015 年全国比赛的冠军中科大队蓝鹰和安徽工业大学的 YUSHAN 队进行了 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如表 1。

表 1

	随机策略	手工代码	学习代码
科大蓝鹰	17.02	13.15	4.15
安徽工业 YUSHN	16.56	10.58	3.54

4 结论与展望

针对 Robocup 防守这一对抗性问题，将多对多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[17]。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，在实际上还有很多改进，比如铲球过程可以用 BP 神经网络优化等。

REFERENCES

[1] Grose B. Collaborative System [J]. AImagazine, 1995, 17(2): 67-85.

[2] Jenninings N R. Commitment, Conventions: the Foundation of Coordination in Multi-Agent system[C]. The knowledge Engining Preview, 1993: 76-241.

[3] Fernando Almeida, Nuno Lau • Luı´s Paulo Reis. An automatic approach to extract goal plans from soccer simulated matches [J]. Soft Comput (2013) 17:835–848.

[4] Jong Yih Kuo, Fu-Chu Huang,* , Shang-Pin Ma, Yong-Yi Fanjiang. Applying hybrid learning approach to RoboCup's strategy [J]. The Journal of Systems and Software 86 (2013) 1933– 1944.

[5] Pedro Henriques Abreu,Daniel Castro Silva, João Portela. Using model-based collaborative filtering techniques to recommend the expected best strategy to defeat a simulated soccer opponent [J]. Intelligent Data Analysis 18 (2014) 973–991 .

[6] Littman M L. Markov games as a framework for multi-agent reinforcement Learning[C]. Proceedings of 11th conference on machine learning, 1994: 103-263.

[7] Chen Mao，KLAUS DORER.Robocup Soccer Server (Users Manual)[Z].2003.

[8] Wellman M P, Hu J L. Multi-agent reinforcement learning: theoretical framework and an algorithm[C]. Processdings of the 15th international conference on machine learning, 1998:242 -250.

[9] Wang Xiaofeng, Sandholm T. Reinforcement learning to play an optimal Nash Equilibrium in team Markov Games[C]. Advances in neural information processing systems 15, Vancouver, 2002: 219-307.Recognition, 1996: 689-693.

[10] Stone P, Veloso M. Multi-agent systems: A survey from a machine learning perspective [J].Autonomous Robots,2000,8(3):345-383.

[11] Weiss G.Multiagent Systems: A modern approach to distributed artificial intelligence [J].the MIT Press,Cambridge,MA,1999.

[12] Durfee E, V R Lesser, Corkill D D. Coorperative Distributed Problem Solving[J].The Handbook of Artificial Intelligence, 1989, 4: 32-37.

[13] Excelente-Toledo C.B，Jennings N.R. The dynamic selection of coordination mechanisms [J].Journal of Autonomous Agents and Multi-Agent Systems,2004,9(1-2):55-85.

[14] Mohen Sharifi, Hamed Mousavian.Predicting the future state of the Robocop simulation envionment:heuristic and neural networks approaches[C]//Proceedings of IEEE International Conference System， Man and Cybernetics. Iran University of Science and Technology, 2003.

[15] Lane D M. Mcfadzean A G. Distributed problem solving and real-time Mechanisms in robot architectures[C]. Engineering Application intelligence,1994,7(2):105 一 117.

[16] Singh M P. Multi-Agent System: A Theoretical Framework for Intentions, Know-how,and Communicatons [M]. Berlin: Springer-Verlag KGS 1994: 1-79.

[17] Wooldrige M, Jennings N R. Intelligent Agents: theory and practice [J].Knowledge Engineering Review,1995,10(2):112-152.

【作者简介】

[1]张俊朋（1989-），男，汉，硕士研究生，研究方向：控制科学与工程。
Email: 2682390094@qq.com

[2]Chen Wei（1953-），female，han，professor，Teseach direction：Control theory and control engineering,Intelligent system。
Email: chenwei@gdut.edu.cn

[3]黄浩辉（1988-），男，汉，硕士研究生,研究方向：控制科学与工程。Email: 1282390094@qq.com