机器学习在RoboCup中的应用研究*)

杨 佩 陈兆乾 陈世福
(南京大学计算机软件国家重点实验室 南京210093)

Research on Machine Learning in the RoboCup

YANG Pei CHEN Zhao-Qian CHEN Shi-Fu
(State Key Lab for Computer Software, Nanjing University, Nanjing 210093)

Abstract RoboCup is a particularly good domain for studying multi-agent systems. A wide variety of MAS issues can be studied in robotic soccer, in which the theory, algorithm and architecture of agent system can be evaluated. Because of the inherent complexity of MAS, there are many interests in using machine learning techniques to handle it. This paper investigates and discusses the machine-learning techniques used in RoboCup. The background is firstly presented and the application of machine learning in RoboCup is lately demonstrated with some top simulation teams. The machine-learning system in NDSocTeam is also introduced. Finally some open issues in this field are pointed out.

Keywords RoboCup, Multi-Agent system, Machine-learning

1. 引言

训练和制造机器人进行足球比赛，是当前人工智能和机器人领域的研究热点之一，创立机器人足球赛(RoboCup)的目的是促进人工智能和机器人学的研究。RoboCup是多agent系统(MAS)研究的理想的测试床，许多MAS中的问题都可以在RoboCup领域中研究，并可以利用它来评价多agent系统的理论、算法和体系结构。鉴于多agent系统固有的复杂性，如状态空间与行为空间十分巨大，利用机器学习(machine-learning)技术来解决这一复杂性有很多益处[1]。机器学习具有一套坚固的机制，能够通过学习经验为agent配备大量行为，这些行为是团队中有效的个人表现以及通过合作完成独立或联合设定的高层目标。许多现存的机器学习技术都可以直接应用到多agent系统中去。

Asada等开发了第一个具有感知能力的机器人球队[2]，这些机器人利用强化学习(reinforcement-learning)训练技术从简单任务中学习将固定的足球踢入球门。这一工作的贡献在于状态行为空间的建立减少了学习任务的复杂性。状态在学习过程中根据每一状态下能采取的最佳动作聚集。另一贡献是将诸如射门和躲避对手这样的低层行为结合起来，这些行为是通过强化学习学习到的[2,3]。

Markov博弈用到的Minimax-Q学习，最先应用在一个抽象仿真足球赛中[4]。这一领域比Soccer Server简单得多，有800个状态，5种动作，没有隐藏信息。每个团队中的队员在网格上移动并且球总是被一个队员控制。利用Minimax-Q学习，队员学习到带球过人的最优策略。

Stone等为一个基于Dynasim仿真团队[5]的球队做机器学习实验。首先使用基于记忆(memory-based)的学习技术使队员学会何时射门何时传球[6]。接着又利用神经网络教会队员将具有不同速度和运行轨迹的球踢进球门内的特定区域[7]。通过在球场的小范围内训练并谨慎选择神经网络的输入表示，agent成功地学习到为它接近运动中的球计，从而能在球场的任何区域内进球。他们也在soccer server系统中使用了相似的技术，将学习到的行为扩展为分层学习系统(layered learning)的一部分[8]。

还有另一个soccer server中的学习实验[9]，队员通过神经网络学习何时射门何时传球。这个实验中，agent在接近对方球门处得到球，守门员挡在前面，而且附近有一个队友。根据足球、守门员和队友的位置，控球队员学习何时最好将球射入球门，何时最好传球。

另外，Balch利用行为种类度量(behavioral diversity measure)进行强化学习框架下的角色学习，并发现对整个团队进行一致强化比对队员进行局部强化有效得多[10]。Luke等利用遗传编程(Genetic Programming)去改善团队的合作行为[11]。

本文的目的在于对机器学习技术在RoboCup中的应用进行研究和探讨。首先用几支世界顶尖球队，CMUnited、ISIS和Karlsruhe Brainstormer，说明了机器学习技术在RoboCup中的具体应用，接着介绍了我们设计的仿真球队NDSocTeam中使用的机器学习系统，最后指出这一领域中一些有待解决的问题。

2. 机器学习在RoboCup中的应用

在此选取世界顶尖球队，CMUnited、ISIS和Karlsruhe Brainstormer，来说明机器学习在RoboCup中的应用。

2.1 CMUnited球队

卡耐基一梅隆大学的Stone等学者针对RoboCup领域的动态、实时和通信不可靠的特点，建立了适合于PTS(Periodic Team Synchronization)领域的agent体系结构，构建了CMUnited队[12]，该队夺取了98年和99年世界机器人足球赛仿真组的冠军，2000年的第三名。他们认为RoboCup是十分复杂的环境：具有有限的通信、实时、环境中有噪声、既有队友

*)基金项目：国家自然科学基金(69905001，60003010)。杨 佩 博士生，主要研究方向为多agent系统、智能agent。陈兆乾 教授，博士生导师，主要研究领域为机器学习、知识工程、神经网络。陈世福 教授，博士生导师，主要研究领域为专家系统、机器学习、神经网络。

又有对手，传感器的输入空间巨大，并存在大量隐藏状态。因此不能学习到传感器输入(sensor input)与激励器输出(actuator output)之间的直接映射。他们提出分层学习法[8]，将问题分解到若干行为层上，并在每一层上使用机器学习技术。学习到的每一层次(子任务)会直接影响下一层的学习。每一层上机器学习方法的选择取决于子任务。利用层次式的任务分解技术，多个机器学习模块可以组合产生更有效的行为。机器学习为训练难度大的行为(如截球和传球评估)搜索数据，或者针对变化的或不可见的环境(传球选择)做出适应性修改。分层学习中任务被分解为三层，截球、传球评估、传球选择。下面将逐一介绍。

2.1.1 *截球(单 agent 行为)* 截球是 agent 的重要技能，由于球的运动有噪声干扰并且 agent 感知能力有限，掌握这一技能并不容易。CMUnited 收集成功截球的实例，并利用有导师的学习技术(supervised learning techniques)生成普遍的截球行为。由于这一技能在比赛过程中不变，此处使用具有4个输入端的神经网络进行离线学习。后卫根据神经网络的输出决定转身角度。在对后卫训练时，通过后卫随机动作并记录下动作的结果来收集训练数据。并学习成功截球的训练实例。

学习的目的是使后卫根据 $BallDist_t$(t 时刻的球距)，$BallAng_t$(t 时刻球的相对角度)，$BallDist_{t-1}$(收到可视信息流时的球距)选择适当的转身角度($TurnAng_t$)。由于 soccer server 不提供明确的球速，而 $BallDist_{t-1}$可以体现球的速度，因而可以根据若干过去的球的位置来估计球速。实际上仅单个过去的球的位置就足以使 agent 学习到截球技能。为了学习到 $TurnAng_t$，该球队使用神经网络，因为它能由连续输入学到连续的输出值。

2.1.2 *传球评估(多 agent 行为)* agent 利用学到的截球技术作为训练多 agent 行为的一部分。当一个 agent 控球并选择向一个特定队友传球时，需要考虑这次传球是否成功。这一评估不仅依赖于队员和对手的位置，也取决于它们接球或截球的能力。因此在为传球评估函数生成训练实例时，使即将接到球的队员和所有对方队员具有上一层学习到的截球能力。考虑到传球评估在所有比赛中基本不变，仍然使用离线学习法：C4.5决策树训练算法[13]。决策树不仅可以忽略无关属性而且可以处理遗漏特征。输入表示共有174种连续属性，一半来自传球者，一半来自接球者。利用5000个训练实例(其中51%成功，42%失败，7%失误)产生具有87个结点的剪枝树。传球者根据决策树提供的可信度评估适合接球的队友。

2.1.3 *传球选择(合作及对立团队行为)* 在这一层，agent 将学到的传球评估能力作为学习传球选择这一团队行为的输入空间。当 agent 控球时，它必须决定将球传给谁。这样的决策依赖于大量信息，包括 agent 在场上的当前位置，所有队友及对手的位置，队友的接球能力，对手的截球能力，队友的决策能力以及对手的策略。考虑到一种决策的优越性只能由团队的长期性能衡量，使用 Q 学习[14]。Q 学习在具有大量输入表示时不理想，然而利用上一层学习到的决策树就可以大大缩小输入空间：只考虑经传球评估后那些可能接到传球的队员。由于学习任务需要在多个 agent 之间分配，并且具有不透明转换(opaque transition)的特性，即，agent 采取一个动作后产生的状态转换是未知的，CMUnited 并没有直接应用 Q 学习，而是创造了一种由 Q 学习得来的多 agent 强化学习方法，TPOT-RL[15]。

和 Q 学习一样，TPOT-RL 学习将状态行为对映射到预期奖赏值上的值函数。但它对传统强化学习算法做出了适应性修改：(1)在多个 agent 之间分配值函数；(2)以递减的探查速率(exploration rate)和递增的开采速率(exploiting rate)同时训练多个 agent；(3)使用行为依赖的特征归纳状态空间；(4)直接从环境中收集长期、折扣后的奖赏值。TPOT-RL 使 agent 团队能够使用极少的训练事例学习有效的策略，甚至在遇到具有大量隐藏状态的巨大的状态空间时也是一样。这是一种在线学习方法。

2.2 ISIS 球队

南加利福尼亚大学的 Tambe 等学者建立了 ISIS 队，该队获得机器人仿真组97年世界季军、98年第四名。ISIS 着重于 agent 设计中的分治(divide-and-conquer)学习法[16]。这是指，运用不同的学习方法对各个 agent 中的技能分别进行学习。目前，它被应用到对射门的学习(利用 C4.5)以及对截球计划的选择(利用强化学习)中。

2.2.1 *射门的离线学习* ISIS 采取自动对射门规则的离线学习。针对3000种射门环境，人类专家为每一种标注最好的射门方向：球门的上、下或中间区域。决策要受其它队员的固定位置和踢球的随机因素(比如，风)的影响。学习系统在1600个随机选择的环境中训练，其它1400个例子用于测试。使用 C4.5作为学习系统，因为它具有合适地表达比赛环境的能力，并能处理遗漏属性和大量训练事例。每个 C4.5训练事例有39个属性。给定由这39个属性定义的比赛环境，决策树选择三个射门方向中最好的那个。最后得到的决策树提供一个射门规则集。这些学习到的规则在 ISIS97中得到了成功应用。

2.2.2 *截球的在线学习* Tambe 等认为截球不是一个简单的静态技能。无论是人类足球或 RoboCup，都有许多影响队员截球的内部和外部环境。例如对手可能更善于奔跑，因此需要队员跑得更快，修改它们截球的路径。球的运动或可视属性也可能显著地影响比赛结果。利用强化学习方法能够使队员在线调整截球技巧，队员在实际比赛中使用 soccer server 提供的感觉信息：球的当前方向和距离，以及方向与距离的变化量。

然而在比赛中并没有许多截球的机会。即使有这样的机会，agent 也常常不能成功截球，例如其它队员恰巧踢开了球，或者球滚出场外等等。为了解决这一问题，使用立即强化(immediate reinforcement)，它在截球计划执行中而不是结束时发生。队员通过顺序执行一系列微计划来截球，每个微计划由一两个快跑之后的转身组成。对微计划中的每一步，ISIS 都期望从 soccer server 获得有关球的位置的任何最新信息。这一期望未达成就造成学习的机会。为了在相似的状态间转移，输入条件聚集。重复发生的错误使一个输入条件对应的微计划改变，尤其是这个输入条件专有的转身增量也要做相应调整。对大部分输入条件来说，实际的转身是根据以下公式由转身增量计算得来的：

$$Turn = BallDir + (TurnInc \times ChangeBallDir)$$

其中，$Turn$ 为实际的转身，$BallDir$ 为球的运动方向，$TurnInc$ 为转身增量，$ChangeBallDir$ 为球运动方向的变化量。

2.3 Karlsruhe Brainstormer 球队

由德国 Karlsruhe 大学的 Riedmiller 等学者创建，该队蝉联机器人仿真组2000年和2001年的世界亚军，并获得2002年的世界第三名。该队的长期目标是将强化学习技术应用在对团队能力的自动学习中。他们认为机器人足球是多 agent 马尔可夫决策问题，任务是利用强化学习找到最优策略[17]。并

将任务分解在两个层次上，技能层和策略层。

2.3.1 技能层 所有的基本技能(截球，带球过人，跑位等)都是用强化学习学习到的。每个技能由基本动作(转身，快跑，踢球)组成。将这个序列称为行动(move)。每个行动持续一定周期并有明确定义的前提和目标。对行动的学习是一个动态优化问题，任务是找到对任何初始状态都具有最小累积代价的策略。在强化学习框架里，建立了以下模型：每个周期里，agent 根据环境选择一个动作，立即代价(immediate cost)大于0。当 agent 到达正最终状态 S^+ 时，命令序列终止并且最终代价值为0。例如，学习截球时，当球在队员踢球的范围之内，命令序列终止，代价为0。然而一些行动也有必须避免的状态。例如，带球过人必须避免球进入对方的踢球范围。这种情况就作为负的最终状态 S^-。当系统遇到这样的状态时，序列终止且代价值为1。该队使用实时动态编程方法[18]来进行强化学习，并用前馈的神经网络估计值函数。

2.3.2 策略层 这一层的任务是进球，对进攻这一团队行为进行了强化学习。策略层的强化学习原则与技能层基本相同，区别在于一个动作不再是基本命令，而是一个行动，即，跑位，截球，传球等。若进球奖赏值为正，否则奖赏值为负。在7个进攻队员对抗7个防守队员和1个守门员的情况下，将所有队员的位置作为一个34维神经网络的输入。通过对导致进球或失球的传球轨迹进行学习，能够找到有助于进球的最短的传球路径。得到球的队员使用这一算法传球，没有得到球的队员对跑位进行学习。为了更新传球轨迹上每一状态的值，使用 TD[1][19]在整个轨迹上传播代价或奖赏值。并使用 RPROP 算法[20]更新神经网络。这一方法在实验中取得了很好的效果[17]。

3. NDSocTeam 中的机器学习系统

RoboCup 是一个复杂动态领域，无法学习到从传感器到激励器的直接映射。因此，我们建立的仿真足球队 NDSocTeam 中进行了层次式的任务分解，采用了分层学习以及多种学习技术相结合的学习方式。这样可以发挥各种机器学习技术本身特有的优势，提高整体学习的效率和效果。在我们设计的球队中，将球员 agent 的学习分为四个层次，如表1所示。

表1 NDSocTeam 中的机器学习系统

层次	学习到的行为	学习方法	训练类型
基本技能层	射门，转身	神经网络、C4.5	离线
复杂技能层	截球、传球、带球过人	强化学习	离线(截球：在线)
合作策略层	变换队形	观察强化学习	在线
对抗策略层	根据对手变换策略	基于示例的学习	离线

下面简单介绍 NDSocTeam 中使用的机器学习系统中每一学习层次的功能。

3.1 基本技能层

底层学习用来训练简单的技能，例如射门，转身等等。由于神经网络能由连续的输入得到连续的输出，我们利用神经网络训练队员进行射门的学习。对于转身，则采取了 C4.5进行学习，因为 C4.5决策树可以有效地处理遗漏属性并忽略无关属性。由于底层涉及的都是一些静态的基本技能，一般不会随着比赛环境的变化而改变，因此都采用了离线学习。

3.2 复杂技能层

这一层则对复杂的技能，尤其是涉及多个球员 agent 之间进行合作的技能进行学习，例如截球、传球以及带球过人等。这些技能均是利用强化学习学习到的。在我们所采用的强化学习中仍然使用了函数估计和强化学习相结合的技术，以处理连续状态空间的问题。除了截球技巧需要队员在线学习外，这类学习过程大多在离线状态下进行。

3.3 合作策略层

在前两层里，我们假设 agent 在固定的队形里，并且不交换位置。但是这个队形在与一个特定对手比赛时并不一定是最好的。我们认为团队可以有多种不同的比赛队形，并且每个队员也可以灵活地交换位置，这样团队的整体性能会更好。在这一层上，我们使用观察(observational)强化学习[21]。队员在比赛中在线调整这一层的合作策略。

3.4 对抗策略层

在现实中，球队会面对各种不同的对手。通过训练球队与许多不同的对手比赛，就可以学到若干不同的策略。这一层我们采用基于示例的学习(case-based learning)进行离线学习。它能够迅速将球队当前的对手与过去曾比赛过的最相似的对手匹配，从而采用以前成功使用过的策略。

4. 进一步研究的问题

4.1 agent 间学习的共享

离线方法适于能在比赛环境之外训练的固定情况。这与人类运动员相似，他们花费大量时间来掌握踢球的基本技能，以求在比赛中自动发挥出来。与人类学习者不同的是，agent 在离线学习时能够与其它队友共享学到的知识。每个独立的 agent 不仅能学到踢球技能，也能将知识传输给队友。CMUnited 就将知识共享方法运用在截球和传球评估层。同时在线学习也能在一定程度上共享。Marcella 等通过实验证明[16]：agent 间学习经验的交流是有益的，然而，不同角色的个体间共享经验或通过使个体执行不同角色来训练个体，都会明显损害团队性能。如何谨慎地进行 agent 间学习的共享，是一个值得研究的问题。

4.2 短期的在线学习和预测

大多数在线学习方法是通过与一个固定的对手多次比赛来学习一种有效的行为。能否在一场仅持续10分钟的比赛中进行有效的在线学习，仍然是一个没有解决的问题，例如，能否迅速学习到一个特定对手 agent 的行为，从而在比赛结束前得到行为预测。

4.3 更多的学习层次

Stone 等提出了从个体行为到团队行为的三个学习层次。在此基础上为其它团队行为和对抗行为建立新的学习层次也是一个有意义的研究课题。

结束语 机器学习是多 agent 系统研究的关键技术，有助于解决多 agent 系统的固有复杂性。而机器人足球领域，具有复杂、动态、实时、不确定性的特点，是机器学习应用的极佳领域。机器人足球赛中的机器学习已经引起了研究者们的广泛关注，近些年来，研究者在这一领域也取得了许多成绩，做出了巨大贡献。

目前我们正在积极地开展对 RoboCup 中机器学习方法的研究[22~24]，建立了分层学习以及多种学习技术相结合的机器学习系统，并将这一研究的初步结果应用到我们研制的仿真球队 NDSocTeam 中。实验表明，NDSocTeam 球队在基本技能的学习和球员之间的协作方面已经达到一定的研制目的，整体的运行效果较好，为进一步对机器学习在 RoboCup

中的应用研究奠定了良好的基础。

参 考 文 献

1 WeiB G, Sen S. Adaptation and Learning in Multiagent Systems. Springer Verlag, Berlin, 1996

2 Asada M, Uchibe E, Noda S, et al. Coordination of multiple behaviors acquired by vision-based reinforcement learning. In: Proc. of IEEE/RSJ/GI Intl. Conf. on Intelligent Robots and Systems (IROS'94), 1994. 917～924

3 Uchibe E, Asada M, Hosoda K. Behavior coordination for a mobile robot using modular reinforcement learning. In: Proc. of IEEE/RSJ/GI Intl. Conf. on Intelligent Robots and Systems (IROS'96), 1996. 1329～1336

4 Littman M L. Markov games as a framework for multi-agent reinforcement learning. In: Proc. of the Eleventh Intl. Conf. on Machine learning, San Mateo, CA, 1994. 157～163

5 http://www.cs.ubc.ca/nest/lci/soccer

6 Stone P, Veloso M. Beating a defender in robotic soccer: Memory-based learning of a continuous function. In: Touretzky D S, Mozer M C, Hasselmo M E, eds. Advances in Neural Information Processing Systems 8, The MIT Press, Cambridge, MA, 1996. 896～902

7 Stone P, Veloso M. Towards collaborative and adversarial learning: A case study in robotic soccer. International Journal of Human-Computer Systems (IJHCS), 1997

8 Stone P, Veloso M. A layered approach to learning client behaviors in the RoboCup soccer server. Applied Artificial Intelligence, 1998, 12:165～188

9 Matsubara H, Noda I, Hiraki K. Learning of coorperative actions in multi-agent systems: A case study of pass play in soccer. In Adaptation, Coevolution and Learning in Multiagent Systems: Paper from the 1996 AAAI Spring Symposium. AAAI Press, Menlo Park, CA, 1996. 63～67

10 Balch T. Learning roles: Behavioral diversity in robot teams: [Technical Report GIT-CC-97-12]. College of Computing, Georgia Institute of Technology, Atlanta, Georgia, March 1997

11 Luke S, Hohn C, Farris J, et al. Co-evolution soccer softbot team coordination with genetic programming. In Kitano H, editor, RoboCup 97: Robot Soccer World Cup I, Springer Verlag, Berlin, 1998. 398～411

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

13 Quinlan J. C4.5 :Programs for Machine Learning. Morgan Kaufman, San Mateo, CA, 1993. 93,95,116,119,122

14 Watkins C. Learning from Delayed Rewards: [PhD thesis]. Kings College, Cambridge, UK, 1989. 93,138

15 Stone P, Veloso M. Team-partitioned, opaque-transition reinforcement learning. In: Asada M, Kitano H, eds. RoboCup-98: Robot Soccer World Cup Ⅰ. Springer Verlag, Berlin, 1999. 135

16 Marsella S, Adibi J, et al. Experiences Acquired in the Design of RoboCup Teams: A Comparison of Two Fielded Teams. Autonomous Agents and Multi-agent Systems, special issue on Best of Agents'99, 1999, 4: 115～129

17 Riedmiller M, Merke A. Karlsruhe Brainstormer-A Reinforcement Learning approach to robotic soccer. RoboCup-01: Robot Soccer World V, LNCS, Springer

18 Barto A G, Bradtke S J, Singh S P. Learning to act using real-time dynamic programming. Artificial Intelligence, 1995, 72: 81～138

19 Sutton R S, Barto A G. Reinforcement Learning: An Introduction. MIT Press, 1998

20 Riedmiller M, Braun H. RPROP: A fast and robust backpropagation learning strategy. In: Jabri M, ed. Fourth Australian Conf. on Neural Networks, Melbourne, 1993. 169～172

21 Andou T. Refinement of soccer agent's positions using reinforcement learning. In: Kitano H, ed. RoboCup-97: Robot Soccer World Cup I, Springer Verlag, Berlin, 1998. 373～388

22 郭磊，艾早阳，陈世福. 多 agent 系统的构造环境 ABE 的研究. 计算机科学，1999，26(8)：58～61

23 葛翔，周志华，陈兆乾. 构造性归纳综述. 计算机科学，1999，26(8)：62～64

24 戈也挺，陈世福，等. 行动推理中若干问题的研究. 计算机科学，2000，27(3)：85～89

(上接第109页)

3 彭聃龄. 语言心理学. 北京：北京师范大学出版社，1991. 2，179，245

4 Groves P M, Schlesinger K. Biological Psychology (2nd Edition). Wm. C. Brown Company Publishers, 1982. 288～292

5 何新贵. 模糊知识处理的理论与技术. 北京：国防工业出版社，1994. 524～551

6 孟绍兰著. 婴儿心理学. 北京：北京大学出版社，1997. 260～300

7 杨雄里著. 视觉的视觉机制. 上海科学技术出版社，1996. 215～267

8 危辉. 智能脑机制的认知结构核心之元态—表象式语义网及其构造：[北京航空航天大学研究生院博士学位论文]. 1998

9 Kimble D P. Biological Psychology, Holt. Rinehart and Winston Inc., 1988. 385～405，179～184

10 Groves P M, Schlesinger K. Biological Psychology (2nd Edition), Wm. C. Brown Company Publishers, 1982. 288～292

11 姚天顺，等. 自然语言理解. 北京：清华大学、广西科学技术出版社，1995. 3

12 危辉. 基于知觉加工模式的发展式分词算法. 计算机研究与发展，2001，38(11)：1281～1289

13 危辉，危炜. 言语获取、理解和生成过程中的语义推理问题. 计算机科学，2002，29(5)：94～96

14 危辉. 讨论：认知语言学对推理型人工智能系统构造的指导作用. 中国人工智能学会第九届年会2001(CAAI-9)，钟义信主编：中国人工智能进展(2001)，北京：北京邮电大学出版社，2001. 1024～1027