基于神经网络的批强化学习在Robocup中的应用

李龙澍[1,2],葛瑞峰[1,2],王慧萍[1,2]
(1.安徽大学 计算机科学与技术学院,安徽 合肥 230039;
2.安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039)

摘 要:模拟机器人足球比赛(Robot World Cup,RoboCup)作为多智能体系统的一个通用的实验平台,通过它可以来检验各种理论、算法和框架等,已经成为人工智能的研究热点。针对在复杂条件下的使用传统Q学习方法所产生的收敛速度缓慢和泛化能力不强的问题,文中使用人工化能力,缩短了学习的时间。并最终将其运用到仿真组比赛的Keepaway模型中,以此验证了该方法的有效性。

关键词:批Q-学习;神经网络;智能体;机器人足球比赛
中图分类号:TP183 **文献标识码**:A **文章编号**:1673-629X(2009)07-0098-04

Application of Batch Reinforcement Learning Based on NN to Robocup

LI Long-shu[1,2], GE Rui-feng[1,2], WANG Hui-ping[1,2]
(1. School of Computer Science and Technology, Anhui University, Hefei 230039, China;
2. Ministry of Education Key Lab. of IC & SP at Anhui University, Hefei 230039, China)

Abstract: As a representative experimental platform of multi-agent system, RoboCup(Robot World Cup) by which various theories, algorithms and architectures can be evaluated, has become the research center of artificial intelligence. For the converge slowly and time consuming problems arised when using the classic Q-learning method in complicated environment, use ANN to represent the Q net and the batch Q learning to process the training data gathered from the environment. By these tactics, improved the generalization capability of the system, and decreased the time cost to learn. It was applied to the experiment of the Keepaway models in the simulation team whose result shows the validity of the method.

Key words: batch Q-learning; neural network; agent; RoboCup

0 引 言

Robocup仿真组作为足球机器人比赛中的一项,无需硬件,提供了一个完全分布式控制、实时异步多智能体的环境,很好地模拟了真实的足球机器人比赛环境,从而为多智能体的智能控制和人工智能理论领域的研究提供了绝佳的实验平台。

在Robocup2D仿真组的比赛中,由于场上状态是一个庞大的、复杂多变的实时系统,并且仿真服务器中具有噪声干扰,所以每个智能体如何在动态变化的环境下选择最优的动作,是一个非常复杂的问题。

Q学习作为强化学习中的一种,可以通过对智能体的训练,使其拥有自主选择最优动作的能力。但是,传统Q函数的实现方法主要是利用表格来表示Q函数,通过查表获得最优Q值[1,2]。当环境的状态集和动作集较大时,Q表需要占用大量的内存空间,而且也不具有泛化能力,这一缺点限制它在连续状态的环境中的应用。

为解决该问题,笔者采用神经网络来拟合Q函数。由于人工神经网络可以逼近任何非线性函数,且具有十分强的自适应、自学习和泛化能力,因此,将人工神经网络与Q学习相结合,可望获得更好的效果。又由于传统的Q学习是在线学习,采集数据较慢;并且在利用最新状态转换信息更新值函数之后就丢弃了,所以采用批Q学习以最大限度地利用智能体在训练中所得到的信息,同时也加快了收敛速度。

1 基于神经网络的批强化学习

1.1 强化学习

强化学习(reinforcement learning)是人工智能中策

收稿日期:2008-11-23;**修回日期**:2009-03-02
基金项目:安徽省自然科学基金(050420204);安徽省高校拔尖人才基金(05025102);安徽省自然科学研究项目(2006KJ098B)
作者简介:李龙澍(1956-),男,教授,博士生导师,研究方向为智能软件和知识工程。

略学习的一种，是一种重要的机器学习方法，又称再励学习、评价学习，是从动物学习、参数扰动自适应控制等理论发展而来。强化学习一词来自于行为心理学，这一理论把行为学习看成是反复试验的过程，从而把动态环境状态映射成相应的动作。该方法不同于监督学习技术那样通过正例、反例来告知采取何种行为，而是通过试错(trial－and－error)的方法来发现最优行为策略[3,4]。

强化学习算法有很多种，例如 TD 算法、Q－学习算法、SARSA 算法 等，其中 Watkin 提出的 Q 学习算法是强化学习算法中最重要的算法之一。文中涉及到的学习方法是 Q－学习。Q－学习是一种无模型强化学习形式。Q－学习是根据状态—动作对(s,a) 映射为期望返回值的动作的价值函数 Q，求解具有不完整信息马尔柯夫行动问题的--种简单方式。

在强化学习模型中处于环境状态 s 的智能体选择并执行一个动作 a，环境状态由于智能体的动作执行迁移到新的状态 s'，同时智能体将会从环境中获得奖赏 r。元组(s,a,r,s') 就构成了一个智能体的经验[5]，智能体就是利用它来改进自身行为的。在经典的强化学习模型中，智能体只使用最新的经验来更新它的策略，即只有最近的状态 － 动作对的价值函数 Q 被更新。例如在 Q 学习中，智能体经历了 t 个经验 $D_t = d_1, d_2, d_3, \cdots, d_t (d_i = (s_i, a_i, r_i, s_i) i = 1,2,3,\cdots,t)$ 在得到经验 d_t 的时候，智能体使用以下的策略来更新 Q 函数。

$$Q_t(s_t,a_t) = Q_{t-1}(s_t,a_t) + \alpha(r_t + \gamma \max_a Q_{t-1}(s_{t+1},a) - Q_{t-1}(s_t,a_t))$$

通过在整个状态空间中反复的迭代，Q 值将最终收敛，即智能体能学习到整体最优的策略。

1.2　基于神经网络的强化学习

在经典的 Q 学习中，Q 值是存储在一张 Q 表里的。Q 表的行和列分别是状态 s 和行为 a，其中的表项就是 $Q(s_i,a_i)$。当将 Q 学习应用在状态空间连续的环境中时，Q 表就会变得非常巨大，不但占用大量内存、查询缓慢，而且也给维护带来了困难。替代的方案是使用函数逼近的方法来存储 Q 值。BP 神经网络具有良好的非线性映射能力，因此很多实验系统中都采用了 BP 神经网络技术，如 Tesauro 的 TD－Gammon 算法[6]。

存储 Q 值的神经网络以状态 s 的各个分量作为输入，输出是在状态 s 下执行动作 $a_i (a_i \in A)$ 所对应的 $Q(s,a_i)$。每次执行一个动作后，Q 值会更新，其 Q 值的变化 ΔQ 为：

$$\Delta Q = \alpha(r + \gamma \max_{a \in A} Q(s',a') - Q(s,a))$$

ΔQ 看作是 BP 网络的输出误差，利用 BP 误差反传算法可以调整权值，从而实现 Q 值的学习。

存储 Q 值的 BP 神经网络训练步骤如下：

1) 初始化，对神经网络的各个节点权值随机赋值；

2) 观察当前状态 s_t；

3) 根据当前状态 s_t，使用神经网络计算每个动作对应的 $Q_t(s_t,a_t)$；

4) 根据 $Q_t(s_t,a_t)$ 选择动作 a_t；

5) 对选定的动作 a_t，使用神经网络计算 Q_{a_t}：$Q_{a_t} = Q_{t-1}(s_t,a_t)$；

6) 执行动作 a_t；

7) 观察新状态 s_{t+1} 和奖励值 r_t；

8) 计算 $Q_{a'} \leftarrow r_t + \gamma \max_{a \in A} Q_{t-1}(s_{t+1},a')$；

9) 通过反传调整网络，使得误差 ΔU 最小：

$$\Delta Q = \begin{cases} Q_{a'} - Q_{a_t} & \text{if } s = s_t, a = a_t \\ 0 & \text{others} \end{cases}$$

10) 返回 2)。

一个 Agent 想得到较大的 Q 值，它在每个状态必须选择具有最高 Q 值的动作，但在学习的初始阶段，Q 值不能准确表示正确的强化值。通常选择具有最高 Q 值的动作会导致 Agent 总是沿着相同路径搜索，那样不可能搜索到较好的值。因此，Agent 选择动作时必须加入随机因素，通常采用的是 Boltzmann 分布[7]：

$$P(s_t,a) = \frac{\exp(\frac{Q(s_t,a)}{T})}{\sum_{b \in A}(\frac{Q(s_t,a)}{T})}, a \in A$$

1.3　批强化学习

在经典的强化学习中，在状态转换后得到的经验被用于值函数的更新之后，就立即丢失了。所以为了使得值函数收敛，就要求智能体在每个状态－动作转换无限频繁的发生。所以要使值函数收敛将花费大量的时间，另外有的经验出现的几率比较小，导致智能体不能充分的学习。

批强化学习模型可以让智能体能从环境中获得经验得到更高效的利用。批强化学习不同于经典的强化学习方法，它将智能体在环境中经历的一系列经验记录下来，再使用经验进行反复的训练。批强化学习的过程简单分为三步：

(1)在 Agent 与环境的交互过程中，记录 m 个情节的经验，并将采集到的经验放到经验集 D 中，转到步骤(2)；

(2) 然后使用 trainBatch 方法利用 D 中的经验进行训练；

(3) 如果 Q 成功收敛，则终止训练，否则清空 D，

转到步骤(1)。

在上述的过程中的 trainBatch 方法有很多种，比如经验重放(Experience Replay)[8]，迭代拟合 Q 方法(Fitted Q Iteration)[8]。在文中使用经验重放方法。它的算法实现如下所示。

经验重放更新 Q 值的算法：

1：$Q \leftarrow Q_0$// 初始化值函数

2：for iteration = 1 to k do// 迭代 K 次

3：for all $i \in [1\cdots \mid D \mid]$ do// 重放每一个经验

4：$Q(s_i,a_i) \leftarrow Q(s_i,a_i) + \alpha(r_i + \gamma \max_{a\in A} Q(s_{i+1}, a) - Q(s_i,a_i))$

5：end for

6：end for

7：Return Q

从上面的算法中可以看到经验重放实际上就是经典强化学习的直接扩展。经典的强化学习在使用当前得到的经验进行一次 Q 值的更新之后，就立刻将此经验丢弃。而经验重放方法是将若干经验存储起来，然后多次使用这些经验对 Q 值进行更新，提高了对训练数据的利用率。同时由于一些经验出现概率比较低，经典的强化学习对于此种经验利用的不充分，导致训练有失全面，而经验重放方法可以有效解决此种问题。

调整后的基于神经网络的批强化学习算法框图如图 1 所示。

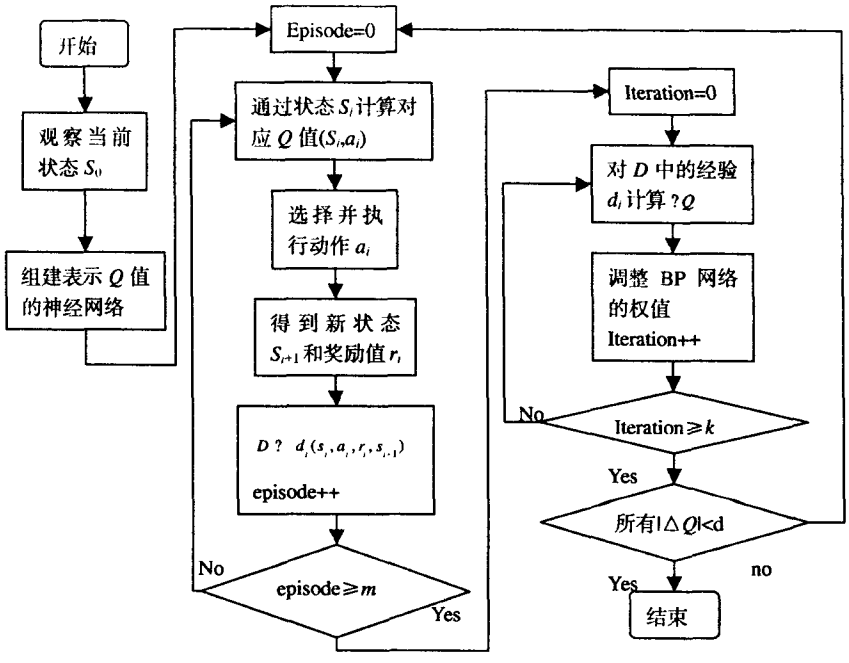

图 1 批强化学习算法框图

2 试验结果

将上述算法应用到了 RoboCup 的 Keepaway (3Vs2)[9]子任务中。Keepaway 子任务是整个任务的简化，由两个队组成。一种称为防守队员(Keeper)，另一种称为进攻球员(Taker)。Keeper 的主要任务就是去尽力守球，使得控球的时间尽可能的长，而 Taker 的主要任务就是从 Keeper 中抢球，使得 Keeper 的控球时间尽可能的短。当球被 Taker 抢走后者球出界，则认为一个情节(episode)结束。整个任务就是由一个个的情节构成的。Keepaway(3Vs2)就是 3 个 Keeper 对 2 个 Taker 的 Keepaway 任务。单单对防守队员进行训练，提高其持球的时间。

沿用 Peterstone 在 Keepaway(3Vs2)研究中使用的由 13 个分量组成的状态 S。其中每一个分量 S_i 是由防守球员和进攻球员之间的距离、角度等等构成的。

动作集合 A 包含持球(HoldBall)，将球传给 K 球员(PassBall(k))，跑到一个最佳等球位置(GetOpen)，跑向球(GoToBall)。这些宏动作都是建立在底层动作基础上的。这样做到目的可以明显减少 Q 学习的状态空间。

在试验中表示 Q 值的神经网络为三层前馈神经网络，隐藏层的单元数是 35 个，网络的各个节点的权值在[-0.5,0.5]范围内随机取值。设置更新 Q 值中的学习率 $\alpha = 10^{-4}$。

图 2 显示的是 500 周期的训练结果，由于使用的是批学习，所以其收敛的速度比经典的 Q 学习快得多。图 3 显示的 4000 个周期的训练结果，可以看到经典 Q 算法到 3500 的时候才达到最好的结果，但是还是比基于神经网络的批 Q 学习的效果要差。因为其使用 Q 表存储 Q 值的时候，要将状态空间离散化，丢失了大量的状态信息，所以其表示的效果要比使用神经网络存储 Q 值差。

3 结束语

针对在 Robocup 仿真组环境下训练多智能体问题，在传统 Q 学习的基础上，采用了 BP 神经网络代替 Q 表，并使用在训练的时候引入了批强化学习的方法，不仅提高了 Q 学习的泛化能力，加速了学习的收敛速度，提高了对训练数据的利用率。并将其用于 Keepaway 实验中，取得了良好的效果，证明了该方法的有效性与可行性。

以后还可以将该方法运用于射门等场合，进一步提高球队的整体作战能力。

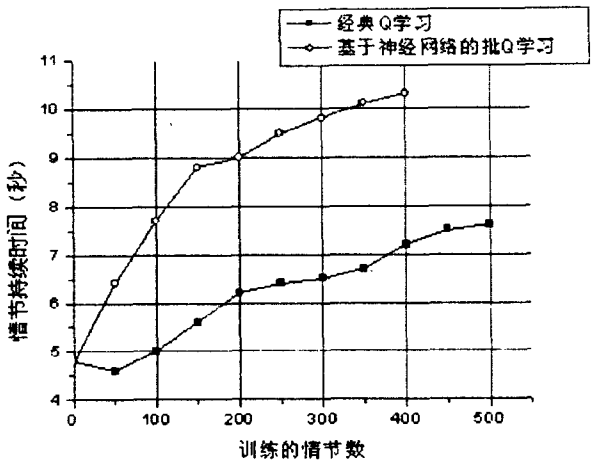

图 2　500 个情节的训练结果

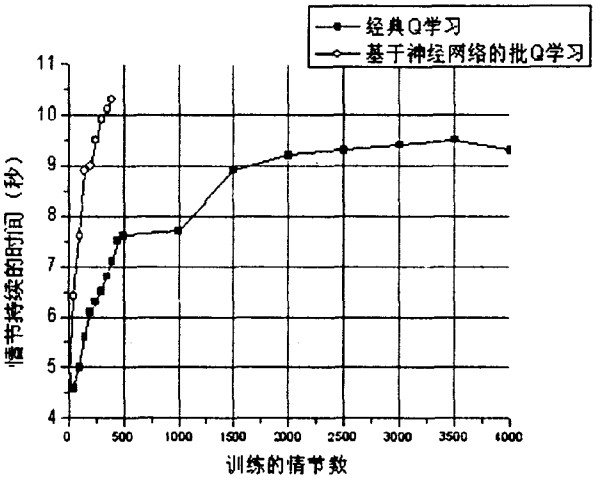

图 3　40000 个情节的训练结果

参考文献:

[1] Kok J R, Vlassis N. Sparse Cooperative Q - learning[C]// Greiner R, Schuurmans D. Proc. of the 21st Int. Conf. on Machine Learning. Banff, Alberta, Canada: ACM, 2004: 481 - 488.

[2] Stone P, Sutton R. Scaling reinforcement learning toward RoboCup soccer[C]// Pro. of the 18th International Conf on Machine Learning. Berkshires, Massachusers: ACM, 2001.

[3] Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey[J]. Journal of Artificial Intelligence, 1996, 4:237 - 285.

[4] Sutton R S, Barto A G. Reinforcement Learning[M]. Cambridge, MA: The MIT Press, 1998.

[5] Lin L J. Self - improving reactive agents based on reinforcement learning, planning and teaching[J]. Machine Learning, 1992, 8:293 - 321.

[6] Tesauro G J. TD - gammon, a self - teaching back gammon program, achieves master - level play[J]. Neural Computation, 1994, 6(2):215 - 219.

[7] 马　勇,李龙澍,李学俊. 基于 Q 学习的 Agent 智能防守策略研究与应用[J]. 计算机技术与发展,2008,18(12):106 - 108.

[8] Ernst D, Geurts P, Wehenkel L. Tree - based batch mode reinforcement learning[J]. J. Mach. Learn. Res., 2005, 6: 503 - 556.

[9] Stone P, Kuhlmann G, Taylor M E, et al. Keepaway soccer: From machine learning testbed to benchmark[C]// RoboCup - 2005: Robot Soccer World Cup IX. New York: Springer - Verlag, 2006:93 - 105.

(上接第 97 页)

HIP 是现有的基于 IPv4 和 IPv6 的互联网下标识符和定位符的分离方案,并在终端间进行认证和建立 IPsec[10]安全关联来提供安全保障。

虽然为了支持 HI 传输层和网络层之间引入了 HIP 层,但是 IETF 协议栈可以继续工作,而且原有的大多数分层网络体系结构不需改变。通过对比分析可以发现 HIP 是更优秀的标识符与定位符的分离方案。

参考文献:

[1] Carpenter B, Crowcroft J, Rekhter Y. IPv4 Address Behaviour Today[S]. RFC2101. 1997.

[2] Meyer D. The Locator/Identifier Separation Protocol[EB/OL]. 2008 - 02 - 27. http://www.ripe.net/ripe/meeting/lisp.

[3] Moskwitz R, Nikander P, Henderson T. Host Identity Protocol[EB/OL]. 2007 - 10 - 30. http://www.ietf.org/draft - ietf - hip - base - 10.txt.

[4] 庄正松,吴家皋,吴清亮,等. 互联网基本服务 IPv4/IPv6 过渡的研究与实现[J]. 计算机技术与发展,2006,16(8):13 - 15.

[5] 徐明伟,吴建平. 主机标识协议研究综述[J]. 小型微型计算机系统,2007,28(2):4 - 5.

[6] 汪文勇. 下一代互联网实名访问机制研究[J]. 电子科技大学学报,2006,35(1):2 - 5.

[7] 徐　剑. 主机标识协议 Puzzle 机制及协议实现研究[DB/OL]. 中国学位论文全文数据库,2006.

[8] 罗利军. 主机标识协议 HIP 的实现—基本通信模块[DB/OL]. 中国学位论文全文数据库,2005.

[9] 刘　欣. 主机标识协议 HIP 的实现—安全模块[DB/OL]. 中国学位论文全文数据库,2005.

[10] 刘淑芝,吴海涛. IPv6 之后的网络安全问题分析[J]. 计算机技术与发展,2006,16(8):243 - 245.