

基于改进的Q学习的RoboCup传球策略研究

周勇, 刘锋

(安徽大学 计算机科学与技术学院, 安徽 合肥 230039)

摘要:模拟机器人足球比赛(Robot World Cup, RoboCup)作为多Agent系统的一个理想的实验平台,已经成为人工智能的研究热点。传统的Q学习已被有效地应用于处理RoboCup中传球策略问题,但是它仅能简单地离散化连续的状态、动作空间。提出将神经网络应用于Q学习,系统只需学习部分状态-动作的Q值即可获得近似连续的Q值,就可以有效地提高泛化能力。然后将改进的Q学习应用于优化传球策略,最后在RoboCup中实现测试了该算法,实验结果表明改进的Q学习在RoboCup传球策略中的应用,可以有效提高传球的成功率。

关键词:RoboCup;神经网络;Q学习;智能体;传球策略

中图分类号:TP242.6

文献标识码:A

文章编号:1673-629X(2008)04-0063-04

Research of RoboCup Pass Strategy Based on Improved Q-Learning

ZHOU Yong, LIU Feng

(School of Computer Science and Engineering, Anhui University, Hefei 230039, China)

Abstract: As the ideal experimental platform of multi-agent system, RoboCup(Robot World Cup) has become the research center of artificial intelligence. Traditional Q-learning dispersed sequential state and action simply on resolving the problem about pass strategy in RoboCup environment. Puts forward a method that neural network is applied to Q-learning, system would output sequential state-action by learning Q-value based on partial state-action and improve generalization ability effectively. The method based on improved Q-learning is proposed to optimize the pass strategy and test the algorithm in RoboCup environment. The experiment shows that improved Q-learning can improve pass efficiency effectively in RoboCup environment.

Key words: RoboCup; neural network; Q-learning; agent; pass strategy

0 引言

机器人足球的最初想法,是由加拿大不列颠哥伦比亚大学的Alan Mackworth教授于1992年正式提出。第一届机器人足球世界杯赛于1997年8月在日本名古屋举行。RoboCup机器人足球赛最重要的目的是检验信息自动化前沿研究,特别是多主体^[1]新成果,交流新思想以及最新进展,从而更好地推动基础研究以及应用基础研究及其成果转化。

RoboCup采用Client/Server方式,由RoboCup联合会提供标准的SoccerServer系统,各参赛队提供各自的Client程序。Client与Server之间通过UDP/IP协议进行通信,Client发送指令控制相应的队员,同时从

Server端接受队员传感器传回的信息。每个Client模块只允许控制一名球员。Server接受双方队伍成员的命令,计算场上所有物体的位置和速度,并向所有队员发送视觉和听觉信息,同时还负责裁判的职责。

对Client来说,传球的准确度对比赛胜负起到了决定性作用。传统的Q学习在处理RoboCup中传球效率不高的问题上,存在着连续的状态、动作空间离散有效性的不足。文中在传统的Q学习基础上,将Q学习与BP神经网络结合,这种模型使得BP网络可以从实际系统学习经验来调整策略,是一个逐渐逼近最优策略的过程,它经过一定周期的学习后用学到的知识训练神经网络,以使得网络逐步收敛到最优状态。并在RoboCup仿真平台上测试,成功地实现了传球策略的优化。

1 Q学习

Q学习是强化学习的一种形式。强化学习(reinforcement learning)是人工智能中策略学习的一种,是一种重要的机器学习方法,又称再励学习、评价学习,

收稿日期:2007-07-05

基金项目:国家自然科学基金(60273043);安徽省自然科学基金(050420204)

作者简介:周勇(1967-),男,安徽合肥人,硕士,讲师,研究方向为机器学习;Agent;刘锋,博士,教授,研究方向为并行分布计算、计算机网络。

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^[2,3]。

如果 Agent 的某个行为策略导致环境正的奖赏(强化信号),那么 Agent 以后产生这个行为策略的趋势便会加强。Agent 的目标可被定义为一个奖赏或回报函数(reward),它对 Agent 从不同状态中选取的不同动作赋予一个数字值,即立即支付(immediate payoff)。Agent 的任务执行一系列动作,观察结果,再学习控制策略,在射门的上层策略中,所希望的控制策略是在任何初始离散状态中选择动作,使 Agent 随时间累积中发现最优策略以使期望的折扣奖赏(回报)和最大。在 Q 学习中 Agent 的某个行为策略导致环境的奖赏用来调整行为策略趋势,基本模型如图 1 所示。

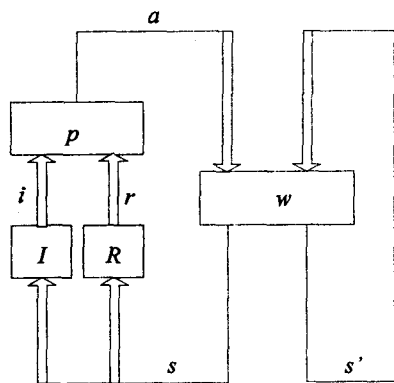


图 1 Q 学习模型

Agent 通过学习一个定义在状态和动作上的数值评估函数,然后以此评估函数的形式实现最优策略将会使过程变得容易。在 Q 学习中把 Q 表示在状态 s 进行 t 动作的预期值: s 是状态向量, a 是动作向量, r 是获得的立即回报, γ 为折算因子。则估计函数 $Q(s, a)$ 被定义为:它的值是从状态 s 开始并使用 a 作为第一个动作时可获得的最大期望折算积累回报。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:

$$P(s_t, a) = \frac{\exp(\frac{Q(s_t, a)}{T})}{\sum_{b \in A} (\frac{Q(s_t, b)}{T})} \quad a \in A \quad (1)$$

用过程描述 Q 学习算法如下:

- (1) 对每个 s, a 初始化表项 $Q(s, a)$;
- (2) 观察当前状态 s , 一直重复做:
 - a) 选择一个动作 a 并执行它;
 - b) 接受到立即回报 r ;
 - c) 观察新状态 s' ;
 - d) 更新 $Q(s, a): Q(s, a) \leftarrow -(1 - \alpha)Q(s, a) + \alpha[r(s, a) + \gamma \text{MAX}_{a'} Q(s', a')]$ ^[4];
 - e) $s \leftarrow s'$, 其中 $\gamma (0 \leq \gamma < 1)$ 是折算因子, 为一常量, α 为学习因子 ($0 < \alpha \leq 1$)。

Q 学习是一个应用很广泛的强化学习算法,但它存在的一定问题。Q 学习算法不能适用于连续状态空间和动作空间的学习,当状态空间很大时, Q 表在内存中的存储以及查询都比较困难。因此将神经网络与 Q 学习相结合来解决 Q 学习中存在的上述问题。

2 基于 BP 神经网络的 Q 学习

人工神经网络是由大量简单的基本元件—神经元相互联结,模拟人的大脑神经处理信息的方式,进行信息并行处理和非线性转换的复杂网络系统^[5,6]。人工神经网络具有良好的自学习、自适应、联想记忆、并行处理和非线性转换的能力,避免了复杂数学推导,在样本缺损和参数漂移情况下,仍能保证稳定输出。

2.1 BP 神经网络

BP 网络模型处理信息的基本原理是:输入信号 X_i 通过中间节点(隐层点)作用于输出节点,经过非线性变换,产生输出信号 Y_k ,网络训练的每个样本包括输入向量 X 和期望输出量 t ,网络输出值 Y 与期望输出值 t 之间的偏差,通过调整输入节点与隐层节点的联接强度取值 W_{ij} 和隐层节点与输出节点之间的联接强度 T_{jk} 以及阈值,使误差沿梯度方向下降,经过反复学习训练,确定与最小误差相对应的网络参数(权值和阈值),训练即告停止。此时经过训练的神经网络即能对类似样本的输入信息,自行处理输出误差最小的经过非线性转换的信息。在模式识别中应用成熟较多的模型是前馈多层式网络中 BP 反向传播模型,也是向前网络的核心部分。模型结构如图 2 所示。

2.2 基于 BP 神经网络和 Q 学习的综合模型

将 Q 学习和神经网络相结合,主要是着眼在利用神经网络的强大存储能力和函数估计能力。对于这一类的学习系统,一般来说神经网络在其中的工作流程是:接收外界环境的完全或不完全状态描述,作为神经网络的输入,并通过神经网络进行计算,输出 Q 学习系统所需的 Q 值或 V 值。采用这种方式可以在较大程度上发挥这两种技术各自特有的优势。其模型结构如图 3 所示。

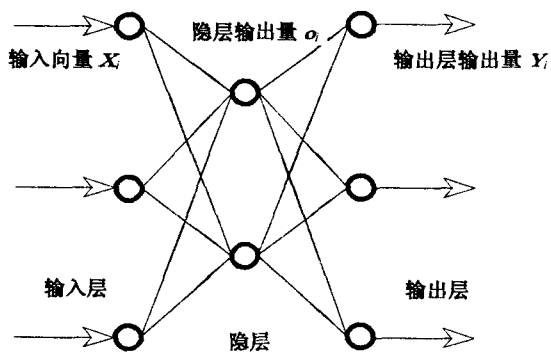


图2 BP网络结构模型

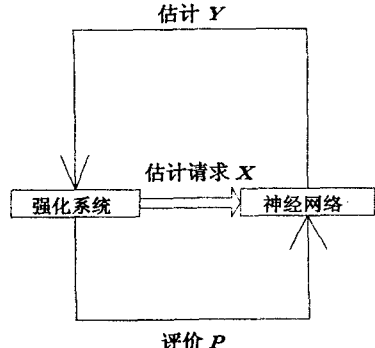


图3 神经网络和Q学习综合模型

集成BP神经网络模型由以下几部分组成:

(1)集成BP神经网络,系统输入 X ,输出 Y 。

(2)随机抖动:产生符合Gaussian分布的抖动量 D ,调整 Y , $Z = Y + D$ 作为对环境的输出,这是强化学习中引入的随机量,以避免陷入局部最小。

(3)评价:环境根据输出 Z 给出评价信号 P ,BP尽量使 P 最大。

(4)延时缓冲:因为强化学习中评价的延迟,而将 D , Y 与评价同步,延时 np 步。

(5)评价选择: $pc = p - pave$; $pave(k+1) = pave(k) + a * (p(k) - pave(k))$,其中 a 是学习率, $pave$ 是系统平均性能的度量, $pave(k)$, $pc(k)$ 表示第 k 次交互时的 $pave$, pc 值。由于延时, $pc(k)$ 与 $X(k - np)$, $Z(k - np)$ 和 $D(k - np)$ 有关。

(6)训练集:令 $Tset(k - np) = Y(k - np)$, $p(k) \leq 0$, $Tset(k - np) = Y(k - np) + D(k - np)$,其它表示当性能提高时接受 $Y + D$ 作为输出,否则以 Y 作为输出, $T(k - np)$ 和 $X(k - np)$ 构成BP网络的训练输入输出对,当训练例达到一定数目时将训练集合提交给网络学习。

3 基于神经网络和Q学习的传球策略

3.1 RoboCup中的传球问题

在Soccer Server中,队员的身体和球都用一个圆来表示,且相互之间位置不允许有重叠部分。当球离

队员的距离(两圆心的距离)小于某个值(称之为传球半径)时,这个队员可以向Server发出一个包括角度和力量两个参数的kick命令,对球施加一个矢量加速度。相对于球员来说,只有当球落在某个范围(称之为控球范围)内时,kick命令才能有效。

在Soccer Server中,球的运动公式如下,其中 v 为速度, P 为位置, t 为周期时刻。

$$(U_x^{t+1}, U_y^{t+1}) = (V_x^t, V_y^t) + (a_x^t, a_y^t) + (r_{rmax}, r_{rmax}) \quad (3)$$

$$(V_x^{t+1}, V_y^{t+1}) = ball_decay * (U_x^{t+1}, U_y^{t+1}) \quad (4)$$

$$(P_x^{t+1}, P_y^{t+1}) = (P_x^t, P_y^t) + (V_x^{t+1}, V_y^{t+1}) \quad (5)$$

文中所说的传球动作决策指的是根据当前周期的初始状态,规划出自本周期起的多个周期内的一系列kick命令,以把球加速到指定的出球速度。在该决策中需要注意的几个影响因素有kick命令对球的加速比、完成踢球动作的周期数、噪声干扰、决策的实时性。

不失一般性,以球的目标速度方向为 x 轴方向,球员的中心点为原点建立坐标系。则每个周期的状态可以用5个量来描述:球的相对位置,球的速度,期望的出球速度大小,队员的速度,队员的身体朝向。其中的主要影响因素是前两者,因此在离线学习中只考虑前两个因素,而把其余的因素也留给了后面的在线规划中去考虑。也就是假设期望出球速度一定,且队员速度为0,身体朝向为0,球的相对位置为 p ,速度为 v 。

关于状态空间 S 的离散,把控球范围离散化成点的集合 $\{P_j\}$ 。同时,把球的速度分布区域离散化为集合 $\{V_j\}$, $j \in [1, 100]$ 。

关于动作空间 A 的离散,把kick指令中的角度以及力量分别进行离散化集合 $\{Ang_i\}$, $\{Power_i\}$, $i \in [1, 100]$ 。

因为球的运动误差与球的速度大小成正比,又有两种类型的噪声干扰,可如下定义代价函数Cost:

$$Cost = 0.1 + 0.01 * V_j + 0.01 * \frac{Abs(|P_j| - player_size_ball - 0.5 * margin)}{margin} \quad (6)$$

其中: $|P_j|$ 表示 P_j 点离原点距离。

Q学习的更新公式如下:

$$Q(S_i, A_i) = (1 - \gamma) * Q(S_i, A_i) + \gamma * MaxQ(S_{i+1}, A_j) + Cost \quad (7)$$

3.2 结合神经网络和Q学习的传球学习过程

神经网络和Q学习结合的算法过程如下:

$$(1) Q(S_i, A_i) = (1 - \gamma) * Q(S_i, A_i) + \gamma * MaxQ(S_{i+1}, A_j) + Cost;$$

$$(2) Y = V_{BP}(X);$$

- (3) $Z = Y + D$;
- (4) $P = P(Z, Q(S_i, A_i))$;
- (5) If $p_c < 0$ then Add(X, Y) Else Add(X, Z);
- (6) If Full(T) then TrainBP(T);
- (7) $S_i = S_{i+1}$, 转步骤(1); 至学习结束。

这里 X 为状态 - 动作对(s, a), V_{BP} 为神经网络输出, $Q(S_i, A_i)$ 为强化学习得到的 Q 值, Z, Y, D 都经过了时延。

4 实验结果及分析

文中的 RoboCup 客户端平台底层代码基于 UVA01, 服务器采用 sim9.3.7 版本。总共进行 1000 个训练周期, 每个训练周期为 5 个服务器周期。实验结果如图 4 所示。

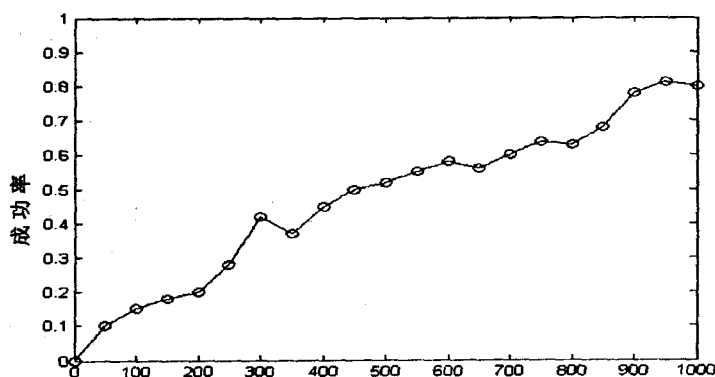


图 4 传球成功率统计曲线

从仿真实验结果来看, 传球的成功率相比以往有了较大的提高, 基于神经网络的强化学习算法在一定程度上提高了传球的成功率。用神经网络来存储 Q 值函数与常见的 Q 表方法相比在时空性能上有了一定程度的提高。首先, 从时间性能上来看, 由于神经网络具有泛化特性, 只需学习部分状态 - 动作的 Q 值, 就

可以获得其它近似状态 - 动作的 Q 值, 因此其学习速度要比 Q 表方法快。其次, 从空间性能上来看, 神经网络所需的存储空间只和网络的连接权个数有关, 而连接权的个数一般要远远小于状态的个数, 所以采用神经网络来存储 Q 值函数的方法所占用的存储也要小于后者。

5 结 语

提出将神经网络应用于 Q 学习, 有效地提高了系统泛化能力, 实现了 RoboCup 中底层传球策略的优化。实验结果说明了与神经网络结合的 Q 学习在一定的训练周期结束后收敛, 最终得到了优化的传球行为序列。今后有关将 RBP 模型应用与上层踢球策略以及改进神经网络的映射能力加强 Q 学习有待进一步研究。

参考文献:

- [1] Stone P. Layered learning in Multi-Agent System [D]. Pittsburgh, PA: Computer Science Department, Carnegie Mellon University, 1998.
- [2] Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey[J]. Journal of Artificial Intelligence, 1996, 4: 237-285.
- [3] Sutton R S, Barto A G. Reinforcement Learning[M]. Cambridge, MA: The MIT Press, 1998.
- [4] Tsitsiklis, John N. Asynchronous stochastic approximation and Q -learning[J]. Machine Learning, 1994, 16(3): 185-202.
- [5] 叶世伟, 史忠植. 神经网络原理[M]. 北京: 机械工业出版社, 2004.
- [6] 丛 爽. 面向 MATLAB 工具箱的神经网络理论与应用[M]. 合肥: 中国科学技术大学出版社, 2003.

(上接第 62 页)

5 结 语

论文的创新之处在于提出了基于 Internet 的教育资源网格体系结构, 采用 RDF 描述教育资源, 并提出索引表和链表结合的方式组织教育资源。链表和索引表结合的资源组织方式, 是根据《基础教育教学资源元数据规范》提出来的, 符合教育资源的特征。同时, 它既适应教育资源网格环境的动态性、异构性, 又能减少查询的成本, 是较好的教育资源组织方式。索引表和链表的动态更新策略, 将是以后研究的重点。

参考文献:

- [1] 朱 莹, 吴军华, 汪婷婷, 等. 网格资源描述技术的比较研

究[J]. 微计算机信息, 2006(22): 176-178.

- [2] Li Juan, Vuong S. A semantics-based routing scheme for grid resource discovery[C]//e-Science and Grid Computing, 2005. First International Conference. [s.l.]: [s.n.], 2005: 438-445.
- [3] CELTS-42 CD1.6. CELTS-42-2002. 教育信息化技术标准[S]. 2002.
- [4] Cai M, Frank M R, Chen J, et al. MAAN: A multi-attribute addressable network for grid information services[C]//Grid Computing, 2003. Proceedings. Fourth International Workshop. [s.l.]: [s.n.], 2003: 184-191.
- [5] Tang C, Dwarkadas S. Hybrid Global-Local Indexing for Efficient Peer-to-Peer Information Retrieval[C]//Proceedings of USENIX NSDI. [s.l.]: [s.n.], 2004.