基于Q学习的Agent智能防守策略研究与应用

马 勇,李龙澍,李学俊
(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;
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系统的一个通用的实验平台,通过它可以来评价各种理论、算法和框架等,已经成为人工智能的研究热点。针对RoboCup仿真中的守门员防守问题,基于Q学习算法,描述了在特定场景中应用Q学习训练守门员的方法和过程。在RobCup中验证了该算法,实现了守门员防守策略的优化。

关键词:Q学习;智能体;机器人足球比赛;防守策略
中图分类号:TP18 **文献标识码**:A **文章编号**:1673-629X(2008)12-0106-03

Research and Application about Defensive Strategy Based on Q Learning

MA Yong,LI Long-shu,LI Xue-jun
(Ministry of Education Key Lab. of IC & SP at Anhui University,Hefei 230039,China;
School of Computer Science and Technology,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. To resolve the problem about defensive strategy of goalie in RoboCup environment, based on Q learning proposed a method which trained goalie. Confirm the algorithm in RoboCup environment and implement the optimization of defensive strategy about goalie.

Key words:Q learning;agent;RoboCup;defensive 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研究重点是球队的高级功能,包括动态多Agent系统中的合作、决策、实时规划和机器学习等当前人工智能的热点问题。因此,在国际人工智能领域,机器人足球被越来越多的人认为是未来50年研究的一个标准问题。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[2],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。Client之间的通讯必须通过SoccerServer来进行。

守门员的防守能力是机器人足球比赛中的一项重要技能。其实质是守门员根据控球队员的移动,改变自身的站位并实时决策站位及防守动作。以往守门员防守策略采取的是手工编码,计算几何防守位置点,因而难以适应机器人足球比赛复杂多变的形势。笔者在这里提出Q学习应用用于守门员的防守策略学习,并对其有效性进行了验证。

1 强化学习

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,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

收稿日期:2008-03-13
基金项目:国家自然科学基金(60273043);安徽省自然科学基金(050420204);安徽省教育厅自然科学研究项目(KJ2007B153)
作者简介:马 勇(1980-),男,安徽和县人,硕士研究生,研究方向为机器学习、智能软件;李龙澍,博导,教授,研究方向为智能软件、知识工程、软件体系结构。

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]。

1.1　强化学习原理及模型

Agent 通过不断尝试错误,从环境中得到奖惩的方法来自主学习到不同状态下哪些动作具有最大的价值,从而发现或逼近能够得到最大奖励的策略。它类似于传统经验中的"吃一堑长一智"。

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(reword),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,我们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。

如图 1 描述:Agent 选择一个动作(action)用于环境,环境(environment)接受动作后状态(state)发生变化,同时产生一个强化信号(reward)反馈给 Agent,Agent 根据强化信号和环境当前状态再选择下一个动作,选择的原则是使受到正强化(奖)的概率增大。

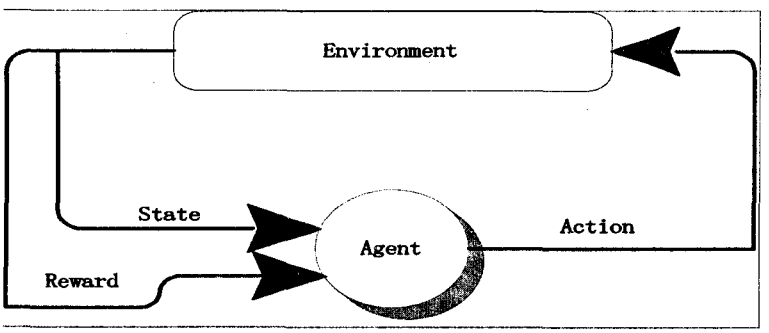

图 1　强化学习的基本模型

1.2　Q 学习算法

Q 学习是强化学习的一种形式,Agent 在任意的环境中直接学习最优策略很困难,因为训练中没有提供 $<s,a>$ 形式的训练样例。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:

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

用过程描述 Q 学习算法如下:

第一步:对每个 s,a 初始化表项 $Q(s,a)$;

第二步:观察当前状态 s,一直重复做:

1) 选择一个动作 a 并执行它;

2) 接受到立即回报 r;

3) 观察新状态 s';

4) 更新 $Q(s,a)$:

$$Q(s,a)<-(1-\alpha)Q(s,a)+\alpha[r(s,a)+\gamma MAXQ(s',a')] \tag{2}$$

5) $s<-s'$,其中 $\gamma(0\leqslant\gamma<1)$ 是折算因子,为一常量。α 为学习因子$(0<\alpha\leqslant 1)$。

2　Q 学习在守门员防守策略中的应用

从上面的强化学习原理中知道,要实现强化学习必须知道状态、动作以及处理这些状态和动作的 Q 函数,在对状态和动作的反复试验当中,还要给出动作的奖赏、设定学习率、折算率等参数。最后要利用 Q 学习算法把这些参数进行最优化,最终得到自己想要的参数值。下面就分析适用于 RoboCup 比赛有关守门员防守策略的状态和 Agent 动作集。

2.1　状态集描述

强化学习应用到 RoboCup 遇到的最大挑战是状态空间的离散化。强化学习适合离散空间求解,而 RoboCup 的环境确是连续的。所以必须离散化环境状态。把 Q 学习所需要的状态集分为守门员的位置区域集、守门员至目标的距离集、守门员的防守角度集等三种集合。

1) 守门员 Agent 的位置区域。

仿真环境中足球场和其中的全部对象都是二维的。任何对象都没有高度的概念。比赛场地的尺寸为 field_length×field_width,缺省值为 105m×68m(单位是没有意义的),球门的宽度为 goal_width,缺省值为 14.64m,是实际的两倍。这里将守门员负责防守的区域离散化,威胁度大的区域,相应的离散值也大。如图

2 所示。

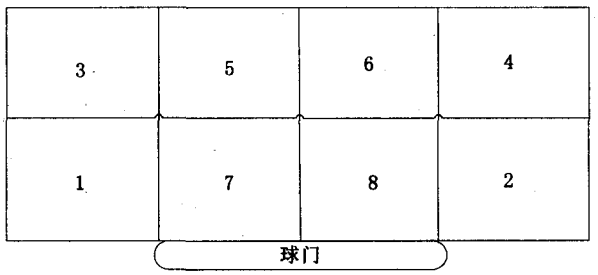

图 2 守门员防守区域离散化点

2) 守门员至控球目标的距离(Dist(A,B))。

球员的视野宽度有三种模式:正常模式[-45,45];宽模式[-90,90],窄模式[-22.5,22.5]。当某个对象在球员的邻域内但在视野之外时,球员只能知道对象的类型(球,其他队员,球门或标志),不知道对象的准确名字。而远处的目标信息则具有不同程度的不确定性。基于守门员特点,选择宽模式视野。距离自己 3 米之内为自己的邻域。定义函数 Dist(A,B),表示 A 至 B 的相对距离,其满足 $0 < \mathrm{Dist}(A,B) \leqslant 15$。

这里将 Dist(A,B) 离散化,在 0 ~ 15 区间内以 3 为单位长度划分为 5 个区间。

3) 守门员防守角度区域。

当没有射门时,守门员的位置选择是基于它所预测的射门路线所确定的。根据球相对于球门的角度,守门员在球门区内选择一个点来防守。球越靠近场地的某一边,守门员当然也应该加强防守这一边。

定义守门员在比赛中的防守角度如图 3 所示。

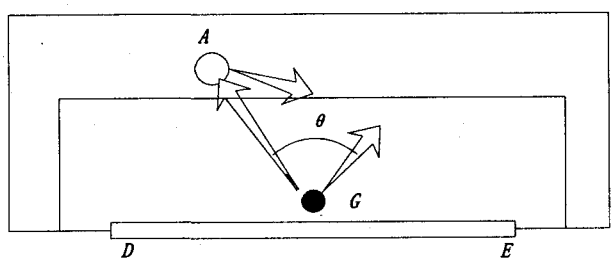

图 3 防守角度示意图

其中 D、E 分别为球门的左右两端点,A 为球,其移动方向如图所示。这里,将守门员的防守角度设为 θ。由 G 作 AE 连线的垂线,垂线与 AG 连线的夹角即为 θ。这里 $\theta \in (0, \pi/2)$。将防守角度按图 2 进行离散化。防守角度分别划分为 $(0,\pi/6]$,$(\pi/6,\pi/3]$,$(\pi/3, 3\pi/2)$ 三个区间。

2.2 动作集描述

球员的基本动作包括 move(移动)、turn(转身)、dash(向前冲)、kick(踢球)、catch(扑球,守门员专用)、say("说"消息)等等。在守门员防守策略中,把动作集进行离散化,在状态 s 下调用自定义的上层动作函数。离散后的动作集为{ SearchBall ,Move,CatchBall,TurnBody},这里的 SearchBall:搜寻目标,Move:移动,CatchBall:捕获球,TurnBody:调整自身状态。

2.3 奖赏的确定

经过上述状态和动作的离散化,守门员的防守学习问题已经转化为一个离散的强化学习问题,现在只需要选择 Q 学习算法,直接以 Q 值作为状态 - 动作对的评估值,进行 $Q(s,a)$ 的强化学习。设计的奖赏规则如下:

1) 如果守门员成功捕获到球,定义回报 $r\mathrm{state} = 5$;否则定义 $r\mathrm{state} = -5$;

2) 定义守门员至球的距离回报

$$r\mathrm{dist} = 5 - \mathrm{deltadist}/10 \tag{3}$$

其中 deltadist 为 2 个周期之间距离差的绝对值。

3) 定义守门员的防守角度回报

$$r\mathrm{angle} = \pi - \mathrm{Angle}/2 \tag{4}$$

这里的 Angle 为图 3 中守门员的防守角度 θ。

这样每个周期球员获得的总回报

$$r = r\mathrm{dist} + r\mathrm{angle} + r\mathrm{state} \tag{5}$$

3 实验结果及分析

文中测试基于 RoboCup 仿真平台。采用 UVA 的代码作为训练球队,重载了原来的守门员防守策略函数。在实际的训练中,设定(1)式中的 $\alpha = 0.1$,$\gamma = 0.95$,训练的场景为 1VS1 对抗。让射门球员分别在图 2 中的 8 个离散化的区域进行射门以训练守门员。总共进行 2000 个训练周期。每个训练周期为 10 个比赛周期。训练结果如图 4 所示。

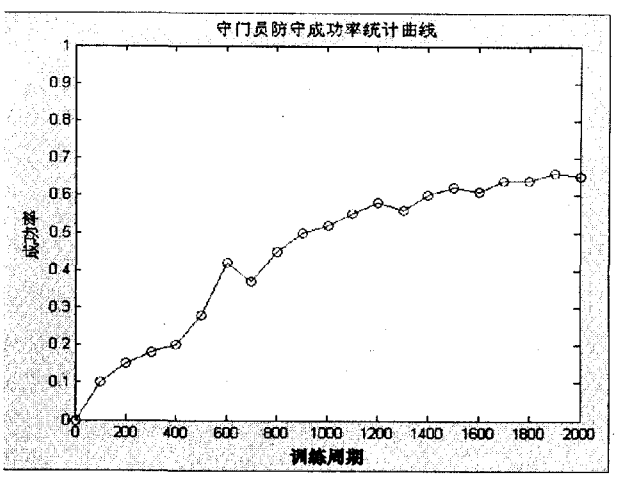

图 4 训练场景与防守成功率关系

用 Q 学习训练前后的守门员程序分别与安徽大学 DreamWing2006、DreamWing2007 队的前锋进行 200 次一对一对抗,防守成功率见表 1。

(下转第 112 页)

一致的。

表 1 对于四分类问题分别采用两种算法的运行结果

类别	预测精度	训练时间	预测时间
二值分类器	83.87%	23.6 秒	0.7 秒
回归机	86.29%	16.4 秒	0.2 秒

根据上面的实验结果和分析可知，基于三值分类器的多分类算法无论在时间复杂度上还是在空间复杂度上都优于基于二值分类器的多分类算法。

4 结束语

在实际运用中，大都需要解决多类别的分类问题，该研究具有较强的实用价值。随着支持向量机在各个领域的广泛运用，如何有效地将该方法推广到多分类问题中已引起越来越多的研究者的重视。该文在对两类分类器分析的基础上，针对 1－v－1SVMs 在分解阶段使用两类分类器造成的信息丢失问题，提出了用回归机作为三值分类器，并将其应用于该阶段。实验证明了该方法的有效性，为支持向量机在多分类问题中的应用起到一定的促进作用。

参考文献：

[1] 邓乃扬，田英杰. 数据挖掘中的新方法——支持向量机[M]. 北京：科学出版社，2004.

[2] Shashua A, Levin A. Taxonomy of large margin principle algorithms for ordinal regression problems[R/OL]. 2007. Leibniz Center for Research, School of Computer Science and Eng. The Hebrew University of Jerusalem. http://citeseer.ist.psu.edu/shashua02taxonomy.html.

[3] Kreel U H G. Pairwise classification and support vector machines[C]// Advances in Kernel Methods: Support Vector Learning. Cambridge, MA, USA: MIT Press, 1999: 255－268.

[4] Angulo C, Gonza'lez L. 1－v－1 tri－class SV machine[C]//In: Proceedings of the 11th European Symposium on Artificial Neural Networks. Bruges, Belgium: [s. n.], 2003: 355－360.

[5] 吴成东，杜崇峰，杨丽英. 基于误差修正码的支持向量机大类别分类方法[J]. 沈阳建筑工程学院学报：自然科学版，2004，20(1)：66－70.

[6] Weston J, Watkins C. Support vector machines for multi－class pattern recognition[EB/OL]. 2007. In Proceedings of 7th European Symposium on Artificial Neural Networks, April 1999. http://citeseer.ist.psu.edu/article/weston99support.html.

[7] 黄 勇，郑春颖，宋忠虎. 多类支持向量机算法综述[J]. 计算技术与自动化，2005，24(4)：61－63.

[8] 唐发明，王仲东，陈锦云. 支持向量机多类分类算法研究[J]. 控制与决策，2005，20(7)：746－749.

[9] Angulo C. Learning with Kernel Machines into a Multi－Class Environment[D]. Spanish: Technical University of Catalonia, 2001.

[10] 刘志刚，李德仁. 支持向量机在多分类问题中的推广[J]. 计算机工程与应用，2004(7)：10－13.

(上接第 108 页)

表 1 200 次对抗成功率对比

	Q 学习前手工编码	Q 学习后手工编码
DreamWing2006	38%	52%
DreamWing2007	32%	48%

可以看出，学习后的守门员防守能力相比学习前有了较大提高。本实验考虑的是一对一场景，学习结果可以用来产生守门员决策，用于防守对方单刀球以及点球。

4 结束语

守门员防守策略问题是 RoboCup 中一个典型的子问题，但它同样是一个实时的 Agent 决策问题。根据我们的经验发现，守门员防守能力作为球队的最后一道防线，往往能在很大程度上影响一支队伍的水平。应用 Q 学习，解决了守门员防守策略问题中状态空间和动作的离散化，实现了 RoboCup 中的守门员防守策略的优化，实验结果表明了 Q 学习在一定的训练周期结束后收敛，最终得到了优化的防守行为策略。研究结果对解决 Agent 智能决策问题具有普遍意义。

参考文献：

[1] Stone P. Layered learning in Multi－Agent System[D]. Pittsburgh: Computer Science Department, Carnegie Mellon University, 1998.

[2] Mihal B, Kay S, Jan W. Learning of kick in artificial soccer[C]//Robot Soccer World Cup IV. Berlin: [s. n.], 2000.

[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] 韩学东，洪炳熔，孟 伟. 强化学习在机器人足球比赛中的应用[J]. 计算机应用研究，2002(6)：79－81.