多智能体的增强学习及其在 RoboCup 中的应用

刘国栋,杨宝庆
LIU Guo-dong,YANG Bao-qing

江南大学 控制科学与工程研究中心,江苏 无锡 214122
School of Communication and Control Engineering,Jiangnan University,Wuxi,Jiangsu 214122,China
E-mail:yibical_2008@yahoo.com.cn

LIU Guo-dong,YANG Bao-qing.Reinforcement learning for Multi-Agents Systems and its application in RoboCup. Computer Engineering and Applications,2008,44(23):46-48.

Abstract: Due to the presence of other agents,the environment of Multi-Agent Systems(MAS) cannot be simply treated as Markov Decision Processes (MDPs).The current reinforcement learning which are based on MDPs must be reformed before it can be applicable to MAS.Based on an agent's independent learning ability,this paper proposes a novel Q-learning algorithm for MAS-an agent learning other agents action policies through observing the joint action.The politicies of other agents are expressed as action probability distribution matrixes.A concise and yet useful updating method for the matrixes is proposed.The full joint probability of distribution matrixes guarantees the learning agent to choose its optimal action.In experiment,the implemention of the agent and the enhancement of AFU shows that the approach is valid and efficient.

Key words: Multi-Agents Systems(MAS);reinforcement learning;Robot World Cup(RoboCup)

摘 要:针对非确定马尔可夫环境下的多智能体系统,提出了多智能体 Q 学习模型和算法。算法中通过对联合动作的统计来学习其它智能体的行为策略,并利用智能体策略向量的全概率分布保证了对联合最优动作的选择。在实验中,成功实现了智能体的决策,提高了 AFU 队的整体的对抗能力,证明了算法的有效性和可行性。

关键词:多智能体;增强学习;机器人世界杯足球锦标赛
DOI:10.3778/j.issn.1002-8331.2008.23.014 **文章编号**:1002-8331(2008)23-0046-03 **文献标识码**:A **中图分类号**:TP78

1 引言

在机器学习范畴内,根据反馈的不同将学习分为监督学习和非监督学习两大类,增强学习属于非监督学习,是一种以反馈为输入的自适应学习方法,通过与环境交互,不断改进最终获得最优行为策略, 由于其在线学习和自适应学习的特点,增强学习是解决智能体策略寻优问题的有效工具,在各个领域获得了广泛地应用[1]。

RoboCup 机器人足球已经成为国际人工智能的一项标准问题。通过学习使足球机器人能够根据环境的变化做出正确的响应。无疑是十分重要的。这被列为了 RoboCup 的挑战之一[2]。

国内外的研究者针对 RoboCup 背景下的学习做了大量工作。德国 Karlsruhe 大学的 Martin 等人对 RoboCup 球队中底层动作进行了学习[3],取得了良好效果;清华大学改进了 Martin 的工作。进一步提高了踢球、带球等基本技能的性能[4];美国德州大学奥斯丁分校的 Peter Stone 针对控球任务,研究了多 Agent 强化学习算法的性能。这些工作主要基于 MDP 框架,应用 Q 学习、Sarsa 等强化算法进行研究,都取得了良好效果,推进了强化学习算法的研究。

但是上述工作。基于 MDP 模型描述问题, 并没有体现多 Agent 这一特征,缺少对多 Agent 行动选择机制的数学刻画。在 RoboCup 中, 多智能体决策这类复杂问题中不能直接引入 Markov 对策模型。必须改进增强学习所依据的环境模型。

本文在传统的 MDP 模型学习算法基础上,考虑到 RoboCup 环境的特点,针对 RoboCup 仿真决策这类复杂问题,提出了学习的基本框架和新的学习算法,并尝试在 RoboCup 仿真球队江大阿福队中应用, 解决了球队中多个智能体的合作决策问题。学习结果在实际比赛中可以直接使用, 具有相当良好的效果,相较过去手工代码的算法,性能有较大提高。验证了 RoboCup 仿真球队中多智能体合作决策学习的可行性和有效性。

2 多智能体 Q 学习

2.1 基于 Markov 对策框架的多智能体强化学习框架

单 AgentMDP(MDP)决策,又称随机对策,可用五元组$<S,\alpha,\{A_i\}_{i\in\alpha},T,\{R_i\}_{i\in\alpha}>$表示。多 AgentMDP(MMDP)是对单 Agent-MDP(MDP)的扩展,也可以用五元组$<S,\alpha,A,T,\{R_i\}_{i\in\alpha}>$来表示。

S 为有限状态集;α 为有限 n 个 Agent 集合;A 为 $A=A_1\times A_2\times\cdots\times A_n$,表示是 n 个 Agent 采取的联合动作$<a_1,a_2,\cdots,a_n>$元素的联合动作空间;T 为 $S\times A_1\times A_2\times\cdots\times A|_\alpha\mapsto PD(S)$,状态转移函数,$PD(S)$表示 S 集合上的状态分布;R_i 为奖赏函数,Agent i

作者简介:刘国栋(1950-),男,教授,硕士生导师,主要研究方向为智能控制、机器人;杨宝庆(1982-),男,研究生,主要研究方向为多智能体。
收稿日期:2007-10-18 **修回日期**:2008-01-21

在状态 S 下,所有 Agent 采取行动后的即时奖赏,$S\times A_1\times A_2\times\cdots\times A|_\alpha\mapsto\rho_0$ 每个 Agent 都以获取最大期望折扣奖赏和为目标。

2.2 多智能体 Q 学习思想

由于多个智能体的存在,不能直接采用单个智能体 Q 学习算法。因学习智能体自身的动作或其他智能体的动作而改变,系统失去了封闭性,后继状态不再仅由当前状态 s 与智能体的动作 a 决定,多智能体增强学习中的回报函数和状态后继函数不能再用 $r(s,a)$和 $S'=\delta(s,a)$来表示。

再者,在多智能体系统中,学习智能体应学习其他智能体的策略,系统当前状态到下一状态的变迁由学习智能体与其他智能体的动作决定,当其他智能体的策略未加时,将造成后继函数的不确定。多数情况下,其它智能体的行为是依据了一定的策略,智能体在某状态下采取的动作是服从一定概率分布的随机行为。因此通过学习过程中对其他智能体的行为进行观察与统计,可学习其它智能体的策略,同时获知该策略对环境的影响,确定回报规则函数和状态后继函数。为此,引入统计方法,通过对状态和动作向量的统计来学习其他智能体的策略。

2.3 多智能体 Q 学习算法

定义学习目标为学习策略 $\pi:S\to A$,S 为有限状态集,A 为智能体动作集合。时刻 t 在状态 S_t 下,智能体选择动作的概率分布表示为 $\pi_t=\{P_1,P_2,\cdots,P_i\}$,策略 $\pi=(\pi_1,\pi_2,\cdots,\pi_t,\cdots)$从状态 S_t开始,按策略 π 获得的期望累计折扣回报 v^π 为:

$$v^\pi(S_t)=E(\sum_{i=0}^{\infty}\gamma^i r_{t+i}) \tag{1}$$

其中,$0\leqslant\gamma\leqslant 1$ 为折扣因子,反映了对当前回报于未来回报的取舍,r_t指每次获得的有界回报,由于是在非确定 Markov 环境下进行学习,累计回报加上期望运算,最佳策略 π^* 是使式(1)获得最大值的策略。

为了描述多个智能体的行为引入动作向量 $\boldsymbol{a}$,对 Q 学习算法[6]改进后有:

$$Q_{(s,\boldsymbol{a})}=E[r_{(s,\boldsymbol{a})}+\gamma V^*(\delta_{(s,\boldsymbol{a})})]=E(r_{(s,\boldsymbol{a})})+\gamma E[V^*(\delta_{(s,\boldsymbol{a})})]=$$

$$E(r_{(s,\boldsymbol{a})})+\gamma\sum_{s'}P_{(s'|s,\boldsymbol{a})}V^*_{(s')} \tag{2}$$

对式(2)进行替换,得:

$$Q_{(s,\boldsymbol{a})}=E(r_{(s,\boldsymbol{a})})+\gamma\sum_{s'}P_{(s'|s,\boldsymbol{a})}\max_{\boldsymbol{a}'}Q_{(s',\boldsymbol{a}')} \tag{3}$$

$P_{(s'|s,\boldsymbol{a})}$表示在状态 s 下,智能体执行联合动作 $a=(a_1,a_2,\cdots,a_i)$后其后继状态为 s'的概率,$\boldsymbol{a}'$为新状态 s'下的联合动作。

用 $\hat{Q}_t$表示第 t 次迭代后 Q 值得近似值,则 Q 值能通过下式进行迭代:

$$\hat{Q}_{t+1(s,\boldsymbol{a})}\leftarrow(1-\alpha_t)\hat{Q}_{t(s,\boldsymbol{a})}+\alpha_t[r_t+\gamma\max_{\boldsymbol{a}'}\hat{Q}^{(s,\boldsymbol{a}')}] \tag{4}$$

α_t是动态学习率。用表示学习智能体的最佳策略,表示学习智能体对智能体 i 在 t 时刻策略的近似估计,将式(4)变为:

$$\hat{Q}_{t+1(s,\boldsymbol{a})}\leftarrow(1-\alpha_t)\hat{Q}_{t(s,\boldsymbol{a})}+\alpha_t[r_t+\gamma\max_{\boldsymbol{a}'}\sum\pi_1^*\prod_{i=2}^{n}\hat{\pi}_t^i\hat{Q}_{t(s,\boldsymbol{a}')}] \tag{5}$$

上式即为提出的多智能体 Q 学习算法,在多智能体环境下智能体可通过该式进行学习。

3 多智能体 Q 学习算法的应用

3.1 在复杂环境中学习的难点与解决方案

对于 RoboCup 仿真环境这类复杂环境,在其中引入学习,最主要的难点在于状态空间极其巨大,同时基本行为对于环境的作用过于基本,通常决策需要建立在基本的行动组合之上。

对于连续的状态空间,引入领域知识,对于关键状态编码,合理地离散化整个状态空间,可以使许多问题的状态空间化简到可解的范围内。

在不同状态下,利用领域知识,组合基本行动为高层动作,产生可选动作集合。此时存在一个新的问题组合而成的高层动作往往执行时间不同,动作执行中可能改变。对此,可以将问题视作 semi-MDP 问题[7],考虑参与对策的所有 Agent 动作的开始与终结时间为计算更新的时刻。

3.2 学习的基本框架与算法

对于复杂问题,需要引入大量的领域知识,将复杂的状态映射到有限状态集合上,同时对具体状态,产生有限的可选动作集合,学习的基本框架如图 1。

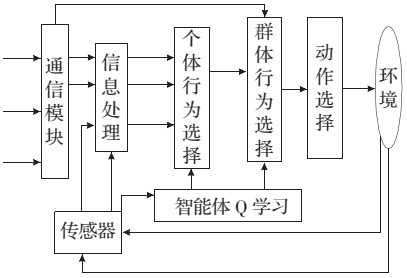

图 1　智能体学习框架

通常的复杂对抗问题中,针对选定的场景(Episode)进行学习,相对于随机布置场景,效果更好。

学习的基本算法如下:

(1)初始化参数。对所有 $s\in S,\boldsymbol{a}\in A^n,\boldsymbol{o}\in O^n,Q_0(s,\boldsymbol{a},\boldsymbol{o})=1,V(s)=1,\pi(s,\boldsymbol{a})=1/|A(s)|,\alpha_0=1.0$,设定场景集合。(2)选择特定场景,初始化实际环境。(3)根据当前状态,确定当前状态下,A^n,O^n的具体含义。(4)行动选择,在当前状态下,以概率 $\pi(s,\boldsymbol{a})$选择行动集 A^n中的 $\boldsymbol{a}$ 行动执行。(5)当观察到对手新动作 $\boldsymbol{o}$ 或自身行动,$\boldsymbol{a}$ 执行结束时,获得即时奖赏 r,当前状态转移到 s'。

修正学习率 $\alpha=\dfrac{\alpha 0}{n(s,\boldsymbol{a},\boldsymbol{o})}$,$n(s,\boldsymbol{a},\boldsymbol{o})$表示$(s,\boldsymbol{a},\boldsymbol{o})$,表示对出现的次数。更新 Q 值:$Q_{t+1}(s,\boldsymbol{a},\boldsymbol{o})=(1-\alpha)\ Q_t(s,\boldsymbol{a},\boldsymbol{o})+\alpha(r+\gamma V(s'))$。求解 $\pi_1^*(s,\cdots)$使得$\max_{\boldsymbol{a}'}\sum\pi_1^*\prod_{i=2}^{n}\hat{\pi}_t^i\hat{Q}_t(s,\boldsymbol{a}')$取最大值,令 V$(s)$等于此最值。

(6)如果当前状态是终止状态(episode end),则转向步骤(2),选择新场景;否则转向步骤(3),继续选择行动。

4 在 RoboCup 决策问题中的应用

RoboCup 仿真环境提供了 Agent 的基本行动,如:kick、dash、turn 等。利用 Peter Stone 提出的 RoboCup 分层学习的思想。在 RoboCup 仿真平台提供的基本动作之上,通过学习或其他方法组合成为高层动作如踢球、截球等。

作为球员的智能体应具有踢球、截球等基本技术,而球队作为一个整体还应具有高层战术策略。高层策略不仅关注个体本身还包括个体间的合作与对抗,如何选择高层策略是一个复

杂的问题。高层策略可看成是智能体在多智能体环境下如何选择最优动作的策略，因此多智能体 Q 学习是解决该问题的有力工具，通过学习智能体可获得高层策略。

在比赛中，引入领域知识，根据具体状态，产生有限可选动作集合。对于多智能体合作与对抗决策问题，进攻队员控球和防守队员控球的行动如下。

进攻队员控球时的行动：*Dribble*(*k*)：带球，控球移动寻找机会，参数 *k* 是可选的带球方位，可选带球方位集合也是利用人类足球知识根据运行时情况确定；*Pass*(*k*)：传球，将球传给其他队员，参数 *k* 是可选的传球方位，可选传球方位集合也是利用人类足球知识根据运行时情况确定；*Shoot*()：射门，直接攻门尝试得。防守队员控球时的行动：*Dribble*(*k*)：带球，控球移动寻找机会，参数 *k* 是可选的带球方位，可选带球方位集合也是利用人类足球知识根据运行时情况确定；*Pass*(*k*)：传球，将球传给其他队员，参数 *k* 是可选的传球方位，可选传球方位集合也是利用人类足球知识根据运行时情况确定；*ClearBall*()：清球，将球踢出界外。

可以给出基本的行为集合：

A={*Shoot*,*Dribble*(1),*Dribble*(2),*Dribble*(3),*Dribble*(4),*Dribble*(5),*Pass*(1),*Pass*(2),*Pass*(3),*Pass*(4),*Pass*(5)}

O={*ClearBall*(),*Dribble*(1),*Dribble*(2),*Dribble*(3),*Dribble*(4),*Dribble*(5),*Pass*(1),*Pass*(2),*Pass*(3),*Pass*(4),*Pass*(5)}

非控球队员的行动则按既定策略进行选择。

在比赛中，球员获球后要根据当前状态执行上面描述的相应动作，描述场上状态采用以下特征来描述：(1) 球员与球坐标；(2)半径 R 内的对手坐标 $op[n][2]$，$R<3$，n 指对手数；(3)半径 L 内的队友坐标 $team[m][2]$，$3<L<25$，m 指队友数；(4)球员与球门两端构成的三角区域内的对手坐标 $opp\ [k][2]$，k 指区域内的对手数。

考虑 3 名进攻队员(7,8,9)与 4 名防守队员(1,2,3,4)的对抗，其中一名进攻队员具有学习能力，学习获球后的动作策略，其余队员为固定策略。进攻队员的目标是成功射门与提高控球时间。训练场景如图 2 所示。

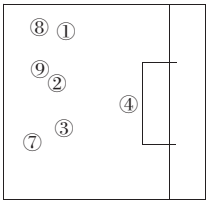

图 2 智能体学习场景

动作评价标准如下：对带球而言，如果带球成功突破对手后且在半径 1.5 m 内没有对手，给其回报 1；对传球而言，如果队友成功获球后且在该队友 1.5 m 半径内没有对手，给其回报 2。学习场景为进攻方从最左方某一位置发球开始至射门成功或球超出以上区域。训练场景通过教练程序设置并记录结果。通过训练，学习智能体可获得最优动作策略。

5 实验结果

实验中，在 RoboCup 仿真平台上，利用离线教练自动布置预先选定的场景，每个场景反复多次运行，每次随机选择预定的 40 个场景之一，每个场景限定执行超时时间为 150 个仿真周期，在实际运行中，每个场景平均的执行时间大约为 50 个仿真周期，即 5 左右(RoboCup 仿真平台上，一个仿真周期时长 100 ms)。在 Intel P4-3.0G 的平台上，经过累计 30 小时的学习，运行约 30 000 个场景后，结果基本稳定。每 200 个场景计算一次学习智能体的动作成功率变化情况如图 3，这里将恰当的带球、传球与射门作为成功动作。

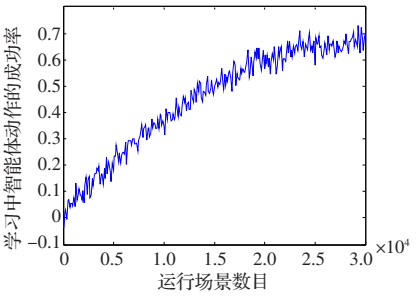

图 3 算法性能曲线

学习后的多智能体程序用于我们设计得 AFU-2007 中与 AFU-2005(采用手工代码)进行了三场比赛并主要对场上的双方的控球率和传球冲功率进行了对比，其结果如表 1 所示。

表 1 算法效果对比

	射门成功率/(%)			控球成功率/(%)		
	1	2	3	1	2	3
AFU-2007	58	61	51	72	78	75
AFU-2005	42	39	49	68	62	66

从表1 中可知 AFU-2007 的无论在控球，还是射门，其能力都是强于 AFU-2005 的，由此可知采用多智能体系统强化学习后，机器人球员的对抗能力有了很大提高。

6 结束语

针对非确定马尔可夫环境的多智能体系统，本文采用一种多智能体系统的 Q 学习框架和算法。该算法通过对联合动作的统计来学习其他智能体的策略，并利用策略概率向量的全概率分布保证了对联合最优动作的选择。同时，该算法将多智能体环境下的学习空间由指数空间降为线性空间，有效提高了学习效率。将该算法应用到多智能体系统 RoboCup 中，实验结果表明了学习算法的有效性。进一步的研究包括提出新的多智能体系统学习算法收敛性的判断；改进学习算法加快学习过程的收敛，使算法更适应在线学习的情况。

参考文献：

[1] Kaelbing L P,Littman M L,Moore A W.Reinforcement learning:a survey[J].Journal of Article Intelligence Research,1996,4(2):237-285.

[2] Kitano H,Tambe M,Stone P,et al.The RoboCup synthetic agent challenge 97[C]//Pollacked M E.Proc IJCAI97.[S.l.]:Morgan Kaufmann,1997.

[3] Riedmiller M,Merke A,Meier D,et al.Karlsruhe brainstormer—a reinforcement learning approach to robotic soccer[C]//Stone P,Balch T,Kraetszchmar G.RoboCup-2000:Robot Soccer World Cup IV.Berlin:Springer Verlag,2001.

[4] Yao Jin-yi,Chen Jiang,Cai Yun-peng,et al.Architecture of tsinghuaeolus[C]//Birk A,Coradeschi S,Tadokoro S.RoboCup 2001:Robot Soccer World Cup V.Heidelberg:Springer-Verlag,2002:491.

[5] Guo Rui,Wu Ming,Peng Jun,et al.A new Q learning algorithm for multi-agent systems[J].Acta Automatic Sinica,2007,33(4):367-372.

[6] Mitchell T M.Machine learning[M].USA:McGraw-Hill Cornpanies Inc,1997.

[7] Puterman M.Markov decision processes[M].New York:Wiley,1994.