基于先验知识的改进强化学习及其在MAS中应用

毛俊杰,刘国栋
MAO Jun-jie,LIU Guo-dong

江南大学 通信与控制工程学院,江苏 无锡 214122
School of Communications and Control Engineering,Jiangnan University,Wuxi,Jiangsu 214122,China
E-mail:maojunjie0802@126.com

MAO Jun-jie,LIU Guo-dong.Modified reinforcement learning based on experience konwledge and its application in MAS.Computer Engineering and Applications,2008,44(24):156-158.

Abstract: In order to increase the speed of the agent learning,which is deficient in the triditional reinforcement learing in MAS. The experience konwledge is used in it,and the conception of intrinsic motivation from psychology is introduced.The intrinsic reinforcement,together with extrinsic reinforcement signal act on the whole process of the learning.At last,this algorithm is used for RoboCup simulation,the results of experiment show that the modified algorithm has faster speed to converge and better performance.
Key words: Multi-Agent System(MAS);experience konwledge;intrinsic motivation;reinforcement learning

摘 要:针对传统的多Agent强化学习算法中,Agent学习效率低的问题,在传统强化学习算法中加入具有经验知识的函数;从心理学角度引入内部激励的概念,并将其作为强化学习的激励信号,与外部激励信号一同作用于强化学习的整个过程。最后将此算法运用到RoboCup仿真中,仿真结果表明该算法的学习效率和收敛速度明显优于传统的强化学习。
关键词:多智能体系统;先验知识;内在激励;强化学习
DOI:10.3778/j.issn.1002-8331.2008.24.047 **文章编号:**1002-8331(2008)24-0156-03 **文献标识码:**A **中图分类号:**TP181

1 引言

近年来,多智能体系统(Multi-Agent System,MAS)的研究已成为计算机科学和人工智能研究的重点。由于多智能体系统的动态性、实时性、分布性、随机性等特点,智能体必须具有学习能力才能与环境自主交互,分析学习外部环境,建立环境模型,模仿人类思维方式学习个体技能、战术策略、协作方式,从而提高多智能体系统的智能水平[1-3]。

在机器学习范畴,根据反馈的不同,学习技术可以分为非监督学习、监督学习和强化学习三大类。其中强化学习是一种以环境反馈作为输入的机械学习方法,主要用于自治Agent如何通过与环境交互来学习最优行为策略。强化学习能够使智能Agent在与环境交互过程中,不断改善自身的性能,以适应不同的环境条件,因此逐渐成为当前机器学习领域研究的一个热点问题[4-5]。

在MAS中,由于外部环境提供信息较少,强化学习的学习效率通常较低[6-7]。针对这个问题,采用具有先验知识的算法来优化学习状态,提高学习效率。

传统的强化学习算法是通过不断地和环境交互,从外界环境中得到评价信号,从而选择合适的动作。而心理学家常常把强化信号区别为内部激励和外部激励[8]。外部激励即从环境得到的奖赏,而内部激励的产生仅仅是因为智能体自我的喜好而与外界无关。内在激励会引导智能个体去尝试各种探索行为、游戏和某些基于好奇而产生的并非由外部奖励的诱惑而产生的行为。心理学家认为,内在激励行为是机体能力提高的重要过程。

本文针对以上两个问题,提出具有先验知识的改进强化学习,首先通过构造经验经验函数,然后从内部以及外部激励两方面考虑,改进强化学习算法,提高学习效率。

2 强化学习

强化学习又称再励学习,是一种以环境的反馈作为输入的、特殊的、适应环境的学习方法。所谓强化学习是指从环境状态到动作映射的学习,以使智能体动作能从环境中获得最大的积累奖赏值。该方法不同于监督学习那样通过正例、反例来告知采取何种行为,而是通过试错来发现最优行为策略,采用统计技术和动态规划方法来迭代逼近在某一环境状态下的动作的效用函数值。

典型的给予折扣回报的强化学习通常可以描述为:

给定 有限状态集合S,有限动作集合A,转移函数$\delta:S\times A\to S'$,回报函数$r:S\times A\to R$。

目标 寻找策略π,使得期望折扣回报总和最大:

$$V(S,\pi)=\sum_{t=0}^{\infty}\gamma E(r_t|\pi,s_t) \quad (1)$$

其中,r_t是t时刻的回报值,$\gamma\in[0,1]$是折扣因子。

作者简介:毛俊杰(1984-),男,硕士研究生,主要研究领域为多智能体,机器学习;刘国栋(1950-),男,教授,博士生导师,主要研究领域为人工智能及应用,机器人系统等。
收稿日期:2007-10-22 **修回日期:**2008-01-22

在任意时刻 t,Agent 处于状态集 S 的某个状态 s_t,根据策略 π 选择一个动作 a_t,经过状态转移函数处理,会使 Agent 处于新的状态 $s_{t+1}=\delta(s_t,a_t)$,同时收到回报 $r_t=r(s_t,a_t)$。函数 $v(s,\pi)$ 表示在状态 s 时策略的期望折扣总回报。

2.1　具有先验知识的强化学习

强化学习的目标是学习在动态环境下如何根据外部评价信号来选择较优动作或者最优动作,本质是一个动态决策的学习过程,但相关环境信息较少,当 Agent 对环境的知识一点也不了解时,它必须通过反复试验的方法来学习,算法的效率不高。

本文提出的基于经验知识的强化学习算法是在标准的强化学习算法中加入具有经验知识的函数 $E:S\times A\rightarrow R$,此函数影响学习过程中智能体动作选择,从而加速算法收敛速度,提高算法的学习效率。

改进算法中的经验函数 $E(s,a)$中记录状态 s 下有关执行动作 a 的相关的经验信息[7]。在算法中加入经验函数的最重要的问题是如何在学习的初始阶段获得经验知识,即如何定义经验函数 $E(s,a)$。这主要取决于算法应用的具体领域,即智能体在与环境交互过程中在线地获得关于环境模型的经验知识。

本文采用的算法将经验函数主要应用在智能体行动选择规则中,动作选择规则如下式:

$$\pi(s_t)=\arg\max_{a_t}[\hat{Q}(s_t,a_t)+\varepsilon E_t(s_t,a_t)] \qquad (2)$$

其中,ε 为一常数,代表经验函数的权重。

2.2　内部激励的强化学习算法

按照经典强化学习的观点,强化学习是一种基于智能体和环境交互的模式,每个时间步,环境都会根据智能体的动作给出一个激励信号,作为智能体的动作的评价。环境给出评价的机制把它叫做评价器[9],如图 1 示。智能体在提高总的时间序列回报的意义下,学习各种动作,并以此改变环境。通过适当的数学假设,智能体的学习问题,被描述为一个马尔可夫决策过程(MDP)的最优解都的估计问题。

但 Sutton 和 Barto 指出[10-11],图 1 这种学习框架是不能和一个生物的学习机制相等同的。一个动物的强化信号的获得,决定于生物体大脑的活动过程,而大脑不仅监视外部信号,同时也受内部状态的影响。真正的评价器应该在大脑中。图 2 把环境细分为外部环境和内部环境,其中内部环境包含了决定回报值的评价器。这个框架仍然包含了回报是外部激励的情况,各种激励被内部环境转化成不同层次的回报。

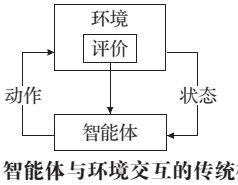

图 1　智能体与环境交互的传统模型

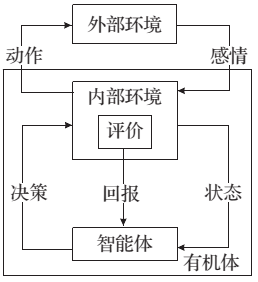

图 2　智能体与环境交互的细分模型

应用强化学习算法一般用定义需要解决问题的回报值的方法来描述问题。比方说赢了一场比赛,回报值为 1,输了就为 0。很多情况下,需要设计巧妙的回报函数。此方法的不同点在于,所谓的内部环境,包含了机体的"动机系统",这个系统应当是一个复杂的系统,而不用对不同的问题重新进行设计。一个专为某个特殊问题而设计的复杂的动机系统显然是没有必要的。

一个智能体的学习过程的结果,应当是积累了一系列可重用的技巧。这里所说的技巧是基于所谓的备选方案(option)理论。备选方案可以看成是一个子程序,它包括策略(环境状态对应的动作)、初始状态集(即备选方案应用条件)、结束条件。它本身并不是一系列的动作,而是一个闭环控制律,也会根据当前的状态改变而改变策略。从理论上讲,当备选方案和基本动作同时作为智能体的策略选择时,通常描述上述行为的马尔可夫决策过程(MDP)已经变成了半马尔可夫过程(SMDPs)。此方法中,备选方案和基本动作一样,在执行过程中的被选择是由概率决定的,而且在与环境的交互中被选择的机率和内部的组成动作的策略都是不断学习更新的。

3　基于先验知识的内部激励强化学习算法

此处,与传统的强化学习不同在于,首先将经验函数用在了智能体的行动选择的策略上,其次学习的智能体大脑中存在一个内部回报学习到的技巧的知识库。

改进后的强化学习原理如图 3 所示。

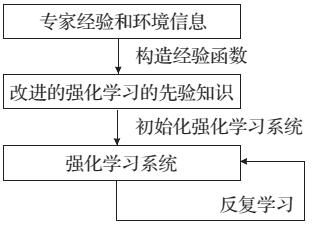

图 3　具有先验知识的强化学习原理

其算法[10]如下:

(1)观察当前状态 s_t,通过下式选择动作 a_t:

$$\pi(s_t)=\arg\max_{a_t}[\hat{Q}(s_t,a_t)+\varepsilon E_t(s_t,a_t)]$$

(2)当前状态 s_t,当前动作 a_t,当前备选方案 o_t,外部回报 r_t^e 和内部回报 r_t^i;

(3)计算下一状态 s_{t+1};

(4)如果 s_{t+1} 是显著事件 e,对 option o_e,若技巧知识库里还还没有,则在知识库里创建 option o_e,把 s_t 加入 o_e 的初始状态集,设置 o_e 的结束概率 $\beta^{o_e}(s_{t+1})=1$,$r_{t+1}^i=\tau[1-P^{o_e}(s_{t+1}|s_t)]$,$\tau$ 是常数。否则,$r_{t+1}^i=0$;

(5)对知识库里所有 $o\neq o_e$,如果 $s_{t+1}\in I^o$ 即 o 的初始状态集,把 s_t 加入 I^o;如果 a_t 是 o 中 s_t 下的贪婪选择的话:

$$P^o(x|s_t)\overset{a}{\leftarrow}[\gamma(1-\beta^o(s_{t+1}))P^o(x|s_{t+1})+\gamma\beta^o(s_{t+1})\delta_{s_{t+1},x}]$$

$$R^o(s_t)\overset{a}{\leftarrow}[r_t^e+\gamma(1-\beta^o(s_{t+1}))R^o(s_{t+1})]$$

$$Q_B(s_t,a_t)\overset{a}{\leftarrow}[r_t^e+r_t^i+\gamma\max_{a\in A\cup Q}Q_B(s_{t+1},a)]$$

(6)$Q_B(s_t,a_t)\overset{a}{\leftarrow}[R^o(s_t)+\sum_{x\in S}P^o(x|s_t)\max_{a\in A\cup Q}Q_B(s_{t+1},a)]$;

(7)对所有 o:

$$Q_B(s_t,o)\overset{a}{\leftarrow}[R^o(s_t)+\sum_{x\in S}P^o(x|s_t)\max_{a\in A\cup Q}Q_B(x,a)]$$

(8)对所有以 s_t 为初始状态集的 option:

$$Q^o(s_t,a_t)\overset{a}{\leftarrow}[r_t^e+\gamma(\beta^o(s_{t+1})\times TVO)+\gamma(1-\beta^o(s_{t+1}))\times\max_{a\in A\cup Q}Q^o(s_{t+1},a)]$$

对每个 $o'\in O,o'\neq o$,若 s_t 也属于 o' 的初始状态集:

$$Q^o(s_t,o')\overset{a}{\leftarrow}R^{o'}(s_t)+\sum_{x\in S}P^o(x|s_t)[\beta^o(x)\times TVO+(1-\beta^o(x))\times\max_{a\in A\cup Q}Q^o(x,a)]$$

其中 TVO 是 optiono 的结束值。

(9)按照 Q_B 以 ε 贪婪策略选择下一步动作 a_{t+1};

(10)对$(s_t,a)\rightarrow s_{t+1}$的转变,赋值 r_{t+1}^e;

(11)$s_t\leftarrow s_{t+1},a_t\leftarrow a_{t+1},r_t^e\leftarrow r_{t+1}^e,r_t^i\leftarrow r_{t+1}^i$,其中,形式 $x\overset{a}{\leftarrow}[y]$是 $x\leftarrow(1-\alpha)x+\alpha[y]$的简写形式。

4 仿真实现结果

本文以 RoboCup 作为实验平台,将提出的算法运用到其中“二对一”这个子问题中,通过此来验证算法是否对提高 MAS 协作能力起到了作用。

“二对一”是 RoboCup 中一个典型的问题,本文假设有两个进攻队员,一个防守球员。进攻方试图通过控球、带球或传球等技术来摆脱对方的抢截或突破防线,如图 4 所示。“二对一”是进攻的两名队员之间相距一定的距离,进行一传一切的基本配合,当突破了防守球员时,任务成功;否则,任务失败。

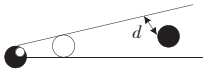

图 4 二对一问题的描述

激励值的设计如下:当采用等待并伺机持球时,有可能延误传球的好时机,故这里把等待后状态的评价作为内部激励,若球被抢走,置-1;若有其他防守队员靠近,则置-0.5;若 d 值变小且队友更利于接球,置 0.5;若几个周期后队友接到球,则外部激励置 1;若球被“二对一”中球员截获,则认为是传球者问题,置外部激励-1;若是被其他球员截获,则认为是接球者问题,置外部激励 0。

以此问题为研究对象,分别采用本文的算法与传统的算法进行训练,并将统计的数据结果作一对比,如图 5 所示,用来说明改进的算法在 MAS 中的有效性。

从对比的结果可以看出,改进后的算法比传统算法无论在学习的速度,还是学习的效果来看,都有了一定的改善。

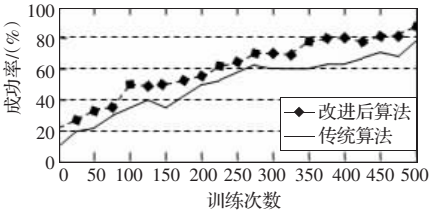

图 5 两种算法训练结果成功率比较图

5 结论

本文针对传统强化学习缺乏环境信息,从而导致学习效率低的问题,采用了基于先验知识的方法,同时,从心理学的角度,将激励行为细分为内部的激励和外部的激励,使 Agent 的学习行为更加智能化,更接近与人类的思考方法和学习方法。将两者结合既在学习方法上进行了优化,提高了学习效率,同时也满足了智能化。

参考文献:

[1] Wang B N,Gao Y,Chen Z Q,et al.LMRL:a multi-agent reinforcement learning model and algorithm[C]//Proceedings of Third International Conference on Information Technology and Applications (ICITA'05),2005.

[2] Piao S H,Hong B R.Fast reinforcement learning approach to cooperative behavior acquisition in multi-agent system[C]//Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems,2002,1:871-875.

[3] 赵丽,董红斌.多 Agent 系统在 RoboCup 中的应用[J].哈尔滨师范大学自然科学学报,2005,21(2):40-45.

[4] Kostas K,Hu H S.Reinforcement learning and co-operation in a simulated multi-agent system[C]//Proceedings of the 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems,1999:990-995.

[5] Yang E F,Gu D B.A multiagent fuzzy policy reinforcement learning algorithm with application to leader-follower robotic systems[C]//Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems,2006:3197-3202.

[6] 杜春侠,高云,张文.多智能体系统中具有先验知识 Q 学习算法[J].清华大学学报:自然科学版,2005,45(7):981-984.

[7] 宋清昆,胡子婴.基于经验知识的 Q-学习算法[J].自动化技术与应用,2006,25(11):10-12.

[8] White R W.Motivation reconsidered:The concept of competence[J].Psychological,Review,1959,66.

[9] Barto A G,Singh S,Chentanez.Intrinsically motivated learning of hierachical collections of skills[C]//Proceedings of the 3rd International Conference on Developmental Learning (ICL'04),LaJolla CA,2004.

[10] 李楠,刘国栋.内在激励强化学习及其在 RoboCup 仿真中的应用[J].计算机仿真,2006,23(4),160-162.

[11] Sutton R S,Barto A G.Reinforcement Learning:an introduction[M].Cambridge,MA:MIT Press,1998.