基于随机对策的团队 CGA 学习

郑延斌,牛丽平
ZHENG Yan-bin,NIU Li-ping

河南师范大学 计算机与信息技术学院,河南 新乡 453007
College of Computer and Information,Henan Normal University,Xinxiang,Henan 453007,China
E-mail:zybcgf@163.com

ZHENG Yan-bin,NIU Li-ping.Research on team CGA learning based on stochastic game.Computer Engineering and Applications,2009,45(23):52-54.

Abstract: In distributed virtual environment,through learning,individual CGA can adapt environment and other CGA in team,so the capability of team problems solving,the adaptability and robust of CGA team have been increased.When the learning based on random games of team CGA has much equilibrium,the equilibrium selection problem of every member in team must be solved.This paper gives a learning method for team CGA called TCCLA.It divides the learning into two levels:manager learning and non manager learning.Every member in team selects its optimization actions according to its preference.Non-manager learns the optimization equilibrium under the direction of manager.So the problem of equilibrium selection has been solved.The IPL algorithm has been improved.The efficiency of TCCLA has been validated through experimentation.
Key words: team CGA(Computer Generated Actor);learn;game;equilibrium;preference

摘 要:分布式虚拟环境中,个体 CGA 通过学习来适应环境和团队中其他 CGA,从而增强团队求解问题的能力,提高团队的适应性和鲁棒性。当基于随机对策的团队 CGA 学习出现多个平衡解时,必须解决平衡的选择问题。提出了一种团队 CGA 学习方法 TCCLA,该方法把团队 CGA 的学习分为两个层次:管理成员的学习和非管理成员的学习,团队中所有成员根据偏好选择最优行为,非管理成员在管理成员的引导下学习到最优平衡,解决了平衡的选择问题,改进了 IPL 算法,实验表明 TCCLA 算法的高效性。
关键词:团队 CGA;学习;对策;平衡;偏好
DOI:10.3778/j.issn.1002-8331.2009.23.015 **文章编号:**1002-8331(2009)23-0052-03 **文献标识码:**A **中图分类号:**TP18

1 引言

学习是 CGA(Computer Generated Actor)根据所积累的经验改善自身行为的能力[1],团队 CGA 学习是团队 CGA 研究的重要内容。团队 CGA 中各个成员通过相互协作来完成团队任务,称为协作团队。协作团队分为完全协作型团队和理性协作型团队。完全协作团队中每个 CGA 具有相同的效用函数,个体 CGA 的最大效用可使整个团队效用的最大化,故可以采用单 CGA 学习算法;理性协作团队中每个成员具有不同的效用函数,个体成员在追求自身效用最大化的同时满足团队效用的最大化,不能采用单 CGA 学习算法,该文研究的是理性协作团队。团队 CGA 学习可以转化为协作对策问题[2],当对策有多个平衡解时,如何确保团队中的每个 CGA 选择同一个平衡解(平衡选择问题),是协作对策必须解决的问题,研究者提出平衡选择方法有:(1)共享经验和通信[3];(2)贝叶斯方法[4];(3)虚拟行动法[5];(4)自适应学习[6];(5)基于社会公约和行为优先顺序的平衡选择方法[7]。这些方法没有充分考虑团队 CGA 的结构,没有体现 CGA 的偏好,并且没有完全体现理想协作团队的特点,该文提出一种团队 CGA 学习算法 TCCLA(Team CGA Cooperation Learning Algorithm),该算法根据 CGA 在团队承担的角色不同,把团队 CGA 学习分为两个层次:管理成员的学习和非管理成员的学习,每个 CGA 根据自身的偏好选择最优行为,非管理成员在管理成员的引导下,选择一致的最优行为,解决了平衡的选择问题。

2 团队 CGA 强化学习算法

2.1 相关定义

在对策解中,Nash 平衡仅仅体现了成员的个体理性,Pareto 平衡仅体现了成员的团队理性,因此 Nash 平衡和 Pareto 平衡都没有很好地体现出协作团队的特点。

定义 1[8] 如果一个 CGA 能够使用 Q-学习算法进行学习,称该 CGA 具有 Q-学习能力。

定义 2[8] 设 Team 为有限个 CGA 组成的团队,满足下面条件的 Team 称为协作团队:(1)Team 中每个成员都具有 Q-学习能力;(2)Team 中成员的联合行动空间中存在一个联合行动 a,

基金项目:河南省科技厅自然科学基金资项目(No.072300410200);河南省教育厅自然科学基金项目(No.2007520027);河南省科技厅重点攻关项目(No.082102210108)
作者简介:郑延斌(1964-),博士,副教授,主要研究领域:虚拟现实、多智能体系统、对策论。
收稿日期:2009-04-28 **修回日期:**2009-06-15

如果 a 使团队中某一个成员的 Q 最大，则团队中所有 CGA 的 Q 值在 a 下都达到最大。

设在某个状态 s 下,团队中一个成员 i,在联合行动 a 下,从环境中得到的奖赏为 $r^i(s,a)$,则相应的 Q 值为:

$$V^i(s)=\max_{a_i\in A_i}\max_{a_{-i}\in A_{-i}} Q^i(s,a_i,a_{-i})$$

相应的修改公式为:

$$Q^i(s,a)=(1-\alpha_t)Q^i(s,a)+\alpha_t[r^i(s,a)+\beta\max_{a_i\in A_i}\max_{a_{-i}\in A_{-i}} Q^i(s',a_i,a_{-i})]$$

状态 s 下的最优 Q 值记为:$Q^*(s,a)$。

定义 3 团队协作平衡 CoopE(Cooperation Equilibrium),一个联合行动 a,在状态 s 下是一个团队协作平衡,如果满足下面的条件:$\forall i\in N,\forall a'\in A,Q_i^*(s,a_i)\geqslant Q_i^*(s,a_i)$。

由定义 3 知道,CoopE 平衡是对策中的一个最优的 Nash 平衡，同时也是最优的 Pareto 平衡,CoopE 平衡实现了对策中 Nash 平衡和 Pareto 平衡的统一。

定理 1 一个协作团队中至少存在一个 CoopE 平衡。

定理 2 如果一个对策有多个 CoopE 平衡，则局中人在每个 CoopE 下的收益是完全相同的。

定义 4 成员 i 对自己行动空间中每个行动的偏好 $p^i(a_l)$,满足:$0\leqslant p^i(a_l)\leqslant 1$,且 $p^i(a_1)+p^i(a_1)+\cdots+p^i(a_m)$,其中 $m=|A_i|$。

定义 5 阶段对策中成员 i 的效用最优值为该成员在 CoopE 下的效用值,记为 $R^i_{\max}(s_t)$。

定义 6 设 a_i 为成员 i 的行动,b 为一个联合行动,则 a_i 与 b 之间的距离定义为:

$$DTA(a_i,b)=\begin{cases}1 & \text{if } a_i=b_i\\ 0 & \text{else}\end{cases}$$

2.2 TCCLA 算法

TCCLA 算法的基本思想是根据团队中每个 CGA 成员承担的角色不同,把团队 CGA 学习分解为两个层:第一层为团队中管理成员的学习,第二层为团队中非管理成员的学习。管理成员通过引导信号来影响其下属成员的行为选择,从而使得所有成员选择相同的 CoopE 平衡。算法模型如图 1 所示。

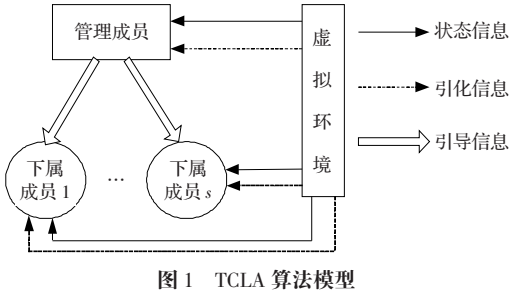

图 1　TCLA 算法模型

2.2.1 团队中管理成员的学习算法

团队中管理成员在每个状态需要维护团队中所有成员的 Q 值表,并根据所有成员的 Q 值表构成的阶段对策求解相应的 CoopE 平衡，根据自己的行动偏好来学习相应的最优行动,并在学习的过程中检查下属成员的行动选择,对没选择相同平衡行为的成员发送引导信息,从而确保所有成员能够学习到同一个 CoopE 平衡。

算法 1 团队中管理成员学习算法

(1)对所有 $s_t\in S,a\in A,i\in N,Q_t^i(s_t,a)=0$。

(2)$p^1(s_t,a_l)\leftarrow 1/|A_1|$,CurrentE←NULL,$t\leftarrow 0$。

(3)在状态 s_t:

①根据自己的行动偏好选择一个行动 a_1;

②观察其他成员的行动:$a_2,\cdots,a_n$,自己及其他成员的奖赏:$r_1,r_2,\cdots,r_n$,及下一个状态 s_{t+1};

③$Q_t^i(s_t,a^t)\leftarrow(1-\alpha_t)Q_t^i(s_t,a^t)+\alpha_t[r_i+\beta\text{CoopE}Q_{t+1}^i(s_{t+1})]$,$i=1,\cdots,n$,其中:$a^t=(a_1,\cdots,a_n)$;

④求对策$[Q^1(s_t)Q^2(s_t)\cdots Q^n(s_t)]$的 CoopE 平衡集合 CE;

⑤求 $R^1_{\max}(s_t)$;

⑥If $r_1\geqslant R^1_{\max}(s_t)$ Then For l=1 to $|A_1|$

If $a_t=a_1$ Then $p^1(s_t,a_l)\leftarrow\min\{1,p^1(s_t,a_l)+\delta(1-p^1(s_t,a_l))\}$

else $p^i(s_t,a_l)\leftarrow\max\{0,p^i(s_t,a_l)-\delta p^i(s_t,a_l)\}$

else if CurrentE=NULL Then

If $\exists b\in CE$ AND DTA(a_1,b)=1 Then CurrentE←b

else $p'(s_t,a_l)\leftarrow\max\{0,p'(s_t,a_l)-\delta p'(s_t,a_l)\}$

else For i=2 to n Do

If DTA(a_i,CurrentE)=1 Then SendCoordMessage(i)

(4)$t\leftarrow t+1$;Goto (3)。

2.2.2 团队中非管理成员的学习算法

算法 2 团队中非管理成员的学习算法

(1)对所有 $s_t\in S,a\in A,Q_t^i(s_t,a)=0$;

(2)$p^i(s_t,a_l)\leftarrow 1/|A_i|$,$t\leftarrow 0$;

(3)在状态 s_t:

①根据自己的行动偏好选择一个行动 a_i;

②观察其他成员的行动 a_{-i},自己的奖赏 r_i,及下一个状态 s_{i+1};

③$Q_t^i(s_t,a^t)\leftarrow(1-\alpha_t)Q_t^i(s_t,a^t)+\alpha_t[r_i+\beta\text{CoopE}Q_{t+1}^i(s_{t+1})]$;

④求 $R^1_{\max}(s_t)$;

⑤If $r_i\geqslant R^i_{\max}(s_t)$ Then For l=1 to $|A_i|$

If $a_i=a_i$ Then $p^i(s_t,a_l)\leftarrow\min\{1,p^i(s_t,a_l)+\delta(1-p^i(s_t,a_l))\}$

else $p^i(s_t,a_l)\leftarrow\max\{0,p^i(s_t,a_l)-\delta p^i(s_t,a_l)\}$

else If HaveCoordMessage(1) Then $p^i(s_t,a_i)\leftarrow\max\{0,p^i(s_t,a_i)-\delta p^i(s_t,a_i)\}$

(4)$t\leftarrow t+1$;Goto (3)。

3 算法分析

(1)算法复杂性分析

①空间复杂度,设 n 为下属 CGA 成员的个数,$|S|$为系统中状态的个数,$|A_i|$为集合 A_i 中包含的行动个数,则算法的空间复杂度为:$n|S||A|^n$。

②时间复杂度，算法的运行时间主要由管理成员在计算

CoopE 平衡时所花费的时间决定,CoopE 平衡可以通过对策矩阵求元素最大值得到,故时间复杂度为:O(n)。

(2)TCCLA 算法特点

TCCLA 算法首先在阶段对策中求 CoopE 平衡,CoopE 平衡是协作团队中最优的 Pareto 平衡,当对策中具有多个 CoopE 平衡时,TCCLA 算法可以保证团队中所有成员快速学习到一个 CoopE 平衡。

TCCLA 算法与 Fulda[9]提出的 IPL(Incremental Policy Learning)算法相似,改进之处为:

(1)IPL 算法是研究重复对策下的协作学习问题,TCCLA算法是随机对策框架下协作学习问题。

(2)IPL 算法没有真正解决平衡的选择问题,是通过相应的条件限制使对策中只有一个最优平衡,TCCLA 算法中当 CGA 选择出现不协调时,团队中管理成员通过引导信息确保选择不一致的成员修改行动,从而使成员能够快速地选择一致。

(3)IPL 算法中,没有考虑当 $r(t)<r_{\max}$ 时的情况,也即对该情况不做处理。这样会出现反复的问题,因为出现这种情况是成员根据自己偏好选择的结果,如果相应的行动偏好不进行修改的话,下次这种结局还会重复出现,从而使学习的速度减慢。

4 仿真实验

为了验证 TCCLA 算法的有效性,用一个 4 个 CGA(坦克 CGA)构成的团队在 10×10 的网格世界中包围一个敌方 CGA(坦克)来进行仿真,如图 2 所示,图中黑色方块代表敌方 CGA,灰色圆圈代表我方 CGA。用我方 CGA 与敌方 CGA 之间的坐标差的和来表示状态,即 $d_i=|x_i-x_p|+|y_i-y_p|$,因此在 10×10 的网格世界中,d 的最大值为 20,由于一个网格中只能有一个 CGA 或者猎物,因此 d 的最小值为 1,当 $d_i=1,i=1,2,3,4$ 时,则任务完成,即团队成功包围敌方(暂不考虑其他情况)。

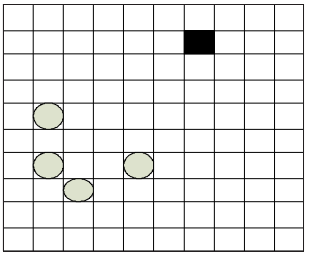

图 2 10×10 网格

如果执行动作的结果使得 CGA 跑出场外或者发生碰撞,则全部弹回原地。如果成功包围敌方则我方每个 CGA 获得的奖赏为 100,否则获得的奖赏为 0。CGA 可执行的动作集合为:{UP,LEFT,RIGHT,DOWN,STOP}。

利用 TCCLA 算法进行了 4 000 次实验,得到了每次实验中 CGA 成功抓获猎物的所需要的步数。设定折扣系数为 β=0.8,学习率为 α=0.9,δ=0.2,TCCLA 算法与虚拟行动进行比较,结果如图 3 所示。图中的数据是每 20 次对相应的数据进行求和平均。从图中可以看出,同等条件下 TCCLA 算法明显优于虚拟行动算法。虚拟行动算法在开始时没有准确得到团队中其他成员的行动策略,因此出现明显的波动,到 1 400 次后算法成功收敛。TCCLA 算法由于 δ 的值对算法效果有一定的影响,开始时出现稍微的波动,在 800 次后就成功收敛。

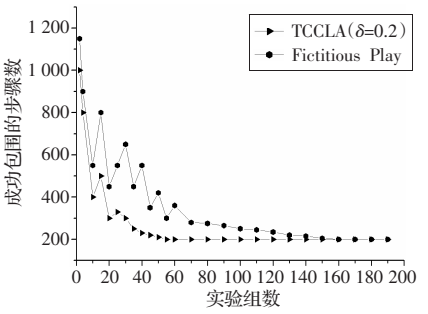

图 3 实验次数与成功包围敌方的次数

图 4 给出了实验中成功抓获猎物的概率与实验组数之间的关系,虽然如图 3 所示成功抓获猎物的步骤数在前期有波动,但是它不影响成功抓获猎物的概率,该图清楚地表明使用 TCCLA 算法成功抓获猎物的概率随着学习次数的增加而逐渐增大的过程。

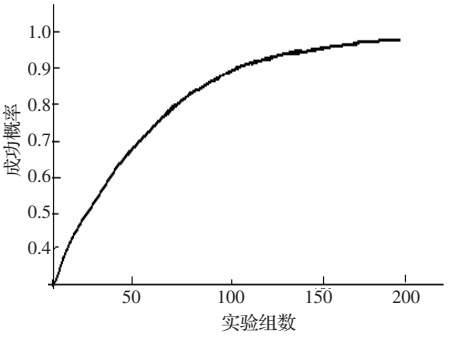

图 4 成功概率随学习过程变换图

5 总结

团队 CGA 协作完成一个复杂任务,从对策论的角度来看,团队中的每个成员既要满足团队理性的要求,同时它又是个体理性的。对策解中 Nash 平衡满足个体理性的要求但是不满足团队理性的要求,Pareto 平衡满足团队理性的要求,但是不满足个体理性的要求。定义了 CoopE 平衡,把 Nash 平衡与 Pareto 平衡结合为一体。CoopE 平衡是对策的最优解,团队 CGA 协作学习就是学习到相应的 CoopE 平衡。为了处理多个 CoopE 平衡的选择问题,给出了一个学习算法 TCCLA。TCCLA 算法是一个成员行动偏好的学习算法。把团队 CGA 学习分为两个层次:管理成员的学习和非管理成员的学习。管理成员在学习的过程中需要维护其下属成员的 Q 值表,并在下属成员的行动选择出现不协调时,引导下属成员的行动选择,使得下属成员能够快速选择同一个 CoopE 平衡。实验结果表明 TCCLA 算法的高效性,TCCLA 算法把 IPL 算法扩展到随机对策框架下,改进了 IPL 算法。

参考文献:

[1] Matatric M J.Reinforcement learning in the multi-robot domain[J]. Autonomous Robots,1997,4(1):73-83.

[2] Littman M L.Markov games as a framework for multi-agent reinforcement learning[C]//Proc of the Eleventh Int Conf on Machine Learning New Brunswich.NJ San Mateo,CA:Morgan Kaufmann Publisher,1994:157-163.

[3] Tan M.Multiagent reinforcement learning:Independent vs. Cooperatives agents[C]//Proceedings of the 10th International Conference on Machine learning.San Francisco:Morgan Kaufmann,1993:487-494.

(下转 103 页)