对称协调博弈问题的多智能体强化学习

王 云,韩 伟
WANG Yun,HAN Wei

南京财经大学 信息工程学院,南京 210046
Information and Engineering College,Nanjing University of Finance and Economics,Nanjing 210046,China
E-mail:dallashw@gmail.com

WANG Yun,HAN Wei.Multi-agents reinforcement learning for symmetrical coordination.Computer Engineering and Applications,2008,44(36):230-233.

Abstract: Considering the problem of robots coordination games,the paper puts forward an agents' belief revision model and a learning algorithm Position-Exchanging Learning(PEL) which is based on the similarity of agents' strategies in coordination games. By position-exchanging,each agent stands from the viewpoint of its opponent and infers opponents' actions.The belief revision model combines the objective observed actions and subjective inferred actions.Coordination is assured by adjusting the belief degree to be 0 or 1.The algorithm PEL is tested in simulations that robots coordinate to avoid collision,and the results prove it performs better than present methods.

Key words: Multi-Agents System(MAS);reinforcement learning;coordination games

摘 要:针对多机器人协调问题,利用协调博弈中智能体策略相似性,提出智能体的高阶信念修正模型和学习方法 PEL,使智能体站在对手角度进行换位推理,进而根据信念修正将客观观察行为和主观信念推理结合起来。证明了信念修正模型的推理置信度只在 0 和 1 两个值上调整即可协调成功。以多机器人避碰为实验背景进行仿真,表明算法比现有方法能够取得更好的协调性能。

关键词:多智能体系统;强化学习;协调博弈
DOI:10.3778/j.issn.1002-8331.2008.36.067 **文章编号:**1002-8331(2008)36-0230-04 **文献标识码:**A **中图分类号:**TP391;TP24

1 引言

机器智能体(以后简写为智能体)的协调问题是多智能体系统(Multi-Agent System,MAS)研究的一个基本课题。诸如多机器人交通控制、避碰、编队、追踪等实际应用问题的研究促使研究者更多地关注以协调或者合作为目标的 MAS。实际上,智能体在群体目标上的合作最终体现在它们行为的协调上。多机器人协调方法大体可以分为两类:一类受社会性昆虫表现出的群体智能启发,依靠生物系统的协调机制来协调智能体之间的行为[1],比如在蚂蚁觅食行为中发现了信息素用以导航[2-3];另一类认为协调是智能体之间相互长期学习和进化的结果,比如狮群捕猎中狮子的占位问题, 狮群并无预先设定的伏击方案,但会根据地形和猎物位置很快"默契"地形成包围。文献[4]提出了多层强化学习算法让机器人学习避碰行为。

本文从博弈学习角度研究机器人协调行为的产生,将机器人每次并发行为看作一次协调博弈。在 MAS 研究中,多智能体的学习是一个具有挑战性的研究内容,在多个智能体共存的环境中,智能体的效用取决于其他智能体的行动策略,因此学习目标不容易被清晰地定义。最近出现的许多研究工作试图将强化学习应用到多个智能体共存的复杂动态环境中,基本的思路是将马尔可夫决策和博弈论结合起来[5-8],使智能体能够学习到对环境和对手的知识。目前关于动态马尔可夫博弈的学习算法主要来自于博弈学习和强化学习两个领域。马尔可夫博弈学习算法除强调收敛性以外,还强调 agent 的个体理性。Min_Max[5]是一个保守的算法,它将博弈的 Nash 均衡作为学习的目标,算法虽然收敛但是智能体的个体理性很弱。对手模型[9]通过建立对手的行为统计模型来最大化自身的期望回报,具有较强的个体理性但不收敛。文献[8]提出的 WoLF(Win or Learn Fast)算法是第一个将收敛性和个体理性结合起来的代表性算法。该算法使用变动的学习率, 智能体首先计算博弈均衡下的效用,若当前效用比均衡状态下的小,则使用较大的学习率,否则使用较小的学习率。可以看出,这个算法本身仍然是以达到博弈均衡状态为目标的,这限制了算法在实际中的应用。在许多应用中, 研究者并不是站在全局角度强调学习算法收敛到 Nash 均衡,而是注重每个智能体是否充分利用了现有知识达到决策最优,这时需要比对手模型(Opponent Model,OM)更强的个体理性。参与博弈的智能体具有相同的环境模型和公共知识,因此在博弈策略上表现出一定相似性。利用这种相似性,智能体可以站在对手角度预测对手的行动,然后结合博弈历史观察到的

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.70802025);江苏省教育厅"青蓝工程"项目;江苏省教育厅自然科学指导计划项目(No.07KJD520070)。
作者简介:王云(1975-),女,讲师,主要研究方向:人工智能,电子商务。
收稿日期:2008-06-17 **修回日期:**2008-10-21

对手行为修正对对手的信念。由于通过推理得到更多信息,算法能够取得更好的协调性能。本文的研究正是从这个角度展开的,与之类似的工作还包括在前期研究中提出的基于高阶信念推理方法的多智能体学习方法[10],但由于计算复杂性等因素,本文智能体只采用一阶推理,并进一步给出了推理模型的参数调整方法。

2　准备知识

2.1　对手模型

Q-学习[11]是单个智能体在复杂动态系统中,通过遍历环境和动作空间,得到最佳行为策略的强化学习算法。若其余 agent 的策略固定,则多智能体环境退化为单智能体环境,采用 Q 学习算法的智能体能收敛到最优策略。

OM 学习是标准 Q 学习算法的一个变型,它假定其余智能体采用一个在状态-行为空间上固定未知的分布,每次博弈观察对手的行为,更新这个分布,然后根据这个分布做出期望值最大意义下的最优决策。这个算法对应于博弈学习理论中的虚拟行动策略[12],该策略用于在可重复剔除最优博弈中寻找 Nash 均衡解。

算法 1　OM 学习算法。

{初始化 $Q(s,a)$为随机数, $\forall s\in S, a_{-i}\in A_{-i}, c(s,a_{-i})\leftarrow 0, n(s)\leftarrow 0$。

repeat

{ 对状态 s,选择 a,使得 $\sum_{a_{-i}}\frac{c(s,a_{-i})}{n(s)}Q(s,(a_i,a_{-i}))$

观察其余智能体的行为 a_{-i},回报 r,下一状态 s'

$Q(s,a)\leftarrow(1-\alpha)Q(s,a)+\alpha(r+\gamma V(s'))$

其中 $V(s)=\max_{a_i}\sum_{a_{-i}}\frac{c(s,a_{-i})}{n(s)}Q(s,(a_i,a_{-i}))$

$c(s,a_{-i})\leftarrow c(s,a_{-i})+1$;

$n(s)\leftarrow n(s)+1$;

}

}

2.2　对称协调博弈

如果博弈所定义的支付矩阵是这样的:在任何均衡点,不仅在给定其他智能体行为的条件下没有智能体愿意改变其行为,而且没有智能体会希望其他智能体改变其行为。这样的问题就是一个协调博弈问题[13]。图 1 是一个协调博弈的示例,为了说明协调博弈与其他非合作博弈的不同,文中给出了它与囚徒困境的对照。在图 1(a)中存在两个纯策略 Nash 均衡:或者两个智能体都选择策略 1 导致支付(6,4),或者两个智能体都采用策略 2 导致支付(6,7)。这两个均衡与图 1(b)中的均衡(坦白,坦白)的区别在于,图 1(b)中的智能体希望他们的对手作出不同的选择。

	A	B
A	6,4	2,0
B	2,0	6,7

图 1(a)　协调博弈

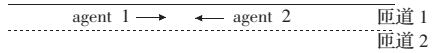
	A	B
A	-1,-1	-8,0
B	0,-8	-5,-5

图 1(b)　囚徒困境

对称协调博弈则是指无角色区分的参与之间进行的协调博弈,它表现在支付函数的对称上。从形式上看:对称协调博弈就是博弈支付矩阵主对角线上的元素都是纳什均衡的博弈。协调博弈的均衡选择并不涉及到回报激励问题而依赖于参与人之间对博弈如何进行有充分相似的信念。正是因为信念形成是一个相对复杂的过程,所以对协调博弈问题的研究也就显得非常复杂,不同的信念形成过程动态就会产生不同的均衡。假定智能体具有支付矩阵完全信息且能观察到其余智能体的行为,则对称协调博弈中的智能体具有了策略相似性(若均衡点位于斜对角线则策略相反)。后文将利用这种相似性(或相反性)提出智能体换位推理学习方法。

2.3　引例及启示

以交通博弈为例考察 OM 的协调性能,结果表明采用 OM 的两个 agent 很难协调成功。举例如下,若 agent 1 初始信念估计(1,1.5),在第一阶段,两个 agent 都认为对方选择匝道 2,因而选择匝道 1。在下一个阶段,信念更新为(2,1.5),两个 agent 都选择匝道 2。结果成为一个交替序列(匝道 1,匝道 1)、(匝道 2,匝道 2)、…,最终无法协调成功。事实上,若对对方行动的先验信念为(a_1,a_2),$|a_1-a_2|<1$,协调就无法成功。



	匝道 1	匝道 2
匝道 1	0,0	1,1
匝道 2	1,1	0,0

图 2　在初始信念分别为(1,1.5),(1,1.2)时虚拟行动策略无法协调成功

交通博弈之所以在虚拟行动下协调不成功,是因为 agent 只是观察对方的行为,对对方的行为进行统计意义上的建模,进而根据自己的理性做出决策。这一假设隐含的前提是:观察到的客观事实代表了对方的主观意图。一般看来,这一过程似乎没有什么问题,但是仔细分析可知,对方 agent 的行为是根据对方对自己的信念改变的,这个信念又来源于自己的行为。也就是说,自身行为是互为因果的,因此对方行为并非是对方意图的完全表达。由此得到启示:agent 不仅要建立对手的行为模型,也要建立对手的行为推理模型,即"知其然",然后"知其所以然"。为此提出换位推理学习方法(Position-Exchanging Learning,PEL),基本思想是:将客观观察行为和主观推理信念区分开来,用换位思考的方式推导出对方行为(称为主观信念行为),给以一定置信度。将客观观察行为和主观信念行为综合起来,建立对手行为的信念修正模型。

3　换位推理学习方法

在马尔可夫博弈$(n,S,A_{1\cdots n},T,\mathrm{R}_{1\cdots n})$中,假定每次博弈后智能体能够观察到其余智能体的回报值,而且每个智能体除维护自身的 Q 表,还对其余的每个智能体分别建立一张 Q 表,即每个智能体都维护 n 张 Q 表,分别用 $Q^{(1)},Q^{(2)},\cdots,Q^{(n)}$表示。每轮博弈之后,$Q$ 表的更新采用非确定环境下的更新规则,即

$$Q_{t+1}^{(i)}(s,(a_i,a_{-i}))\leftarrow(1-\alpha_{t+1})Q_t^{(i)}(s,<a_i,a_{-i}>)+\alpha_{t+1}[r_i(s,<a_i,a_{-i}>)+\gamma V(s')](i=1,\cdots,n)$$

其中 $V(s')=\max_{a_i}\sum_{a_{-i}}p(a_{-i}|s')Q(s',<a_i',a_{-i}>)$,$\alpha_{t+1}=\frac{1}{1+visits_{t+1}(s,a)}$。$visits_{t+1}(s,a)$为 $t+1$ 次更新中状态-动作对$<s,a>$被访问的总次数。

定义 1　函数 $B_i^o(a_{-i}):A_{-i}\rightarrow R^+$,称为客观信念修正函数。其中 $A_{-i}=A_1\times A_2\times\cdots\times A_{i-1}\times A_{i+1}\times A_n$,$B_i^{o,t}(a_{-i})=B_i^{o,t-1}(a_{-i})+\begin{cases}1,\text{若 } a_{-i}^{t-1}=a_{-i}^{t}\\0,\text{其他}\end{cases}$,

$P_i^{o,t}(a_{-i})=\dfrac{B_i^{o,t}(a_{-i})}{\sum_{a^{-i}} B_i^{o,t}(a_{-i})}$为对手行动的联合概率。

对客观概念修正的一个改进方法是强调最近观察的对手行动,称为指数加权改进。

定义 2 函数 $B_i^{o'}(a_{-i}):A_{-i}\to R^+$,称为客观信念修正函数。其中

$A_{-i}=A_1\times A_2\times\cdots\times A_{i-1}\times A_{i+1}\times A_n$

$B_i^{o',t}(a_{-i})=\beta B_i^{o',t-1}(a_{-i})+\begin{cases}1,若\ a_{-i}^{t-1}=a_{-i}^{t}\\0,其他\end{cases}$，$\beta\in(0,1]$

定义 3 若智能体 i 站在智能体 j 的角度，引用智能体 j 的 Q 表 $Q^{(j)}$ 采用 OM 学习方法,进行一次博弈,计算应该采取的行动 $a_j^{(1)}$,称这一行动为对智能体 j 的主观信念行动。

定义 4 若智能体 i 以经过客观信念修正后,对智能体 j 的主观信念行动为 $a_{i,j}$,则信念修正函数定义为:

$P_{ij}^s:A_j\to[0,1],P_{ij}^{s,t}(a_{jk})=P_{ij}^{o,t}(a_{jk})+\delta_{ij}I(a_{jk})$

其中,$P_{ij}^{o,t}(a_{jk})$是联合概率 $P_i^{o,t}(a_{-i}):A_{-i}\to[0,1]$对 j 的边际分布。$I(a_{jk})$为标号函数,$I(a_{jk})=\begin{cases}1,若\ a_{jp}=a_{i,j}\\0,否则\end{cases}$;$\delta_{ij}\in[0,1]$为推理置信度，表征智能体 i 对智能体 j 行为进行推理的信任度。δ_{ij}越大,算法就更多地体现为推理和预测,反之,算法更注重从博弈历史中进行强化学习。特别的,对所有其余智能体 $j(j\neq i)$,$\delta_{ij}=1$ 时算法退化成推理算法,$\delta_{ij}=0$ 时算法退化为一般的 OM 学习算法。

例 1 以 2.3 节提到的交通博弈为例，取 $\beta=0.5,\delta_{12}=0.4,\delta_{21}=0.1$。智能体 1 初始信念为(1,1.5),第一次协调不成功后,客观信念更新后为(1.5,0.75),转化为概率为(0.67,0.33)。由于推测得到的主观信念行动为匝道 2,信念修正后为(0.67,0.73),从而选择匝道 1。智能体 2 初始信念为(1,1.2),客观信念更新后为(1.5,0.6),转化为概率为(0.71,0.29)。由于推测得到的主观信念行动为匝道 2,信念修正后为(0.568,0.432),从而选择匝道 2。协调成功。

表 1　智能体 1 客观、主观信念修正及动作选择

客观更新的信念	(1,1.5)	(1.5,0.75)	(0.75,1.38)
OM 下的概率	(0.4,0.6)	(0.67,0.33)	(0.35,0.65)
OM 下的行为	匝道 1	匝道 2	匝道 1
对手的信念	(0.5,0.5)	(1.25,0.25)	(1.63,0.13)
主观信念行动		匝道 2	匝道 2
信念修正		(0.67,0.73)	(0.35,1.05)
PEL 下的行动		匝道 1	匝道 1

表 2　智能体 2 客观、主观信念修正及动作选择

客观更新的信念	(1,1.2)	(1.5,0.6)	(1.75,0.3)
OM 下的概率	(0.46,0.54)	(0.71,0.29)	(0.85,0.15)
OM 下的行为	匝道 1	匝道 2	匝道 2
对手的信念	(0.5,0.5)	(1.25,0.25)	(0.63,1.13)
主观信念行动		匝道 2	匝道 1
信念修正		(0.71,0.39)	(0.95,0.15)
PEl 下的行动		匝道 2	匝道 2

算法 2 换位推理学习方法。

步骤 1 观察对方行为,根据定义 2 进行客观信念修正。

步骤 2 计算其余各个智能体的边际客观信念。

步骤 3 分别对每个对手计算主观信念行动。

步骤 4 根据定义 4 对每个对手进行主观信念更新。

步骤 5 由每个智能体主观信念，导出对所有对手的联合信念。

步骤 6 根据 OM 方法做出行为决策，即选择对联合信念最大的行为组合的最优反应。

4　算法分析

例 1 中,可以观察到:对智能体 1,有 $0<\dfrac{\alpha_1^{(t)}}{\beta_1^{(t)}}<1$,且 $\dfrac{\alpha_1^{(t)}}{\beta_1^{(t)}}$随 t 增大而减小,因此智能体 1 将一直选择匝道 1;对智能体 2,有 $\dfrac{\alpha_2^{(t)}}{\beta_2^{(t)}}>1$,且 $\dfrac{\alpha_2^{(t)}}{\beta_2^{(t)}}$随 t 增大而增大,因此智能体 2 将一直选择匝道 2。也就是说,若第一次信念修正能够协调成功,则智能体相互信念出现分化,并且这种分化随着博弈进行不断加强。

定理 1 若 PEL 方法在第 t 次博弈协调成功后,保持 $\delta_{ij}(i,j=1,2,\cdots,n$ 且 $i\neq j)$不变,则 PEL 在第 t 次之后的博弈中一直协调成功。

证明 记智能体的 t 次博弈后信念为$(\alpha_i^{(t)},\beta_i^{(t)})(i=1,2,\cdots,n)$,考察智能体 i 和智能体 j 博弈的情况。若第 t 次博弈协调成功,则$(\alpha_i^{(t)}-\beta_i^{(t)})(\alpha_j^{(t)}-\beta_j^{(t)})<0$。不妨设$\dfrac{\alpha_i^{(t)}}{\beta_i^{(t)}}<1,\dfrac{\alpha_j^{(t)}}{\beta_j^{(t)}}>1$。只需证明 $\dfrac{\alpha_i^{(t+1)}}{\beta_i^{(t+1)}}<1,\dfrac{\alpha_j^{(t+1)}}{\beta_j^{(t+1)}}>1$。令第 t 次博弈后,智能体 i 的客观信念为$(O_{i\alpha}^{(t)},O_{i\beta}^{(t)})$,转化为概率为$(p_{i\alpha}^{(t)},p_{i\beta}^{(t)})$,因为第 t 次协调成功,则 $O_{i\alpha}^{(t+1)}=\beta O_{i\alpha}^{(t)},O_{i\beta}^{(t+1)}=\beta O_{i\beta}^{(t)}+1$。因此

$$p_{i\alpha}^{(t+1)}=\frac{O_{i\alpha}^{(t+1)}}{O_{i\alpha}^{(t+1)}+O_{i\beta}^{(t+1)}}=\frac{\beta O_{i\alpha}^{(t)}}{\beta O_{i\alpha}^{(t)}+\beta O_{i\beta}^{(t)}+1}<\frac{O_{i\alpha}^{(t)}}{O_{i\alpha}^{(t)}+O_{i\beta}^{(t)}}=p_{i\alpha}^{(t)}$$

由 $p_{i\alpha}^{(t+1)}+p_{i\beta}^{(t+1)}=1$ 得到 $p_{i\beta}^{(t+1)}>p_{i\beta}^{(t)}$，因此$\dfrac{\alpha_i^{(t+1)}}{\beta_i^{(t+1)}}=\dfrac{p_{i\alpha}^{(t+1)}}{p_{i\beta}^{(t+1)}+\delta_{ij}}<\dfrac{p_{i\alpha}^{(t)}}{p_{i\beta}^{(t)}+\delta_{ij}}=\dfrac{\alpha_i^{(t)}}{\beta_i^{(t)}}<1$ 成立。同理可分析智能体 j 的情况。证毕。

定理 1 说明第 t 次协调博弈成功可以保证第 t 次以后都成功，因此问题转化为如何使第 t 次博弈中的智能体协调成功。关于推理置信度与协调性能的关系,有如下定理。

定理 2 通过调整 $\delta_{ij}\in\{0,1\}$,PEL 方法必定能协调成功。

证明 考虑协调均衡点位于支付矩阵的斜对角线的情况。假定第 t 次因智能体 i,j 均选取行为 A 而协调不成功，记智能体的 t 次博弈后 OM 下的概率为$(p_{i\alpha}^{(t)},p_{i\beta}^{(t)})$、$(p_{j\alpha}^{(t)},p_{j\beta}^{(t)})$,且有$(p_{i\alpha}^{(t)}-p_{i\beta}^{(t)})(p_{j\alpha}^{(t)}-p_{j\beta}^{(t)})>0$，即 OM 下仍然协调不成功。设智能体 i,j 在 PEL 下的信念分别为$(p_{i\alpha}^{(t)},p_{i\beta}^{(t)}+\delta_{ij})$、$(p_{j\alpha}^{(t)},p_{j\beta}^{(t)}+\delta_{ji})$,若 PEL 协调成功,则智能体 i 和智能体 j 有且仅有一个改变了其行为,即$(p_{i\alpha}^{(t)}-p_{i\beta}^{(t)})(p_{i\alpha}^{(t)}-p_{i\beta}^{(t)}-\delta_{ij})<0$ 或者$(p_{j\alpha}^{(t)}-p_{j\beta}^{(t)})(p_{j\alpha}^{(t)}-p_{j\beta}^{(t)}-\delta_{ji})<0$,综合

这两种情况,得到

$$(p_{i\alpha}^{(t)}-p_{i\beta}^{(t)})(p_{i\alpha}^{(t)}-p_{i\beta}^{(t)}-\delta_{ij})(p_{j\alpha}^{(t)}-p_{j\beta}^{(t)})(p_{j\alpha}^{(t)}-p_{j\beta}^{(t)}-\delta_{ji})<0$$

记,$d_i^{(t)}=p_{i\alpha}^{(t)}-p_{i\beta}^{(t)}$,$d_j^{(t)}=p_{j\alpha}^{(t)}-p_{j\beta}^{(t)}$,上式变为:

$$d_i^{(t)}\delta_{ji}+d_j^{(t)}\delta_{ij}>d_i^{(t)}d_j^{(t)}+\delta_{ji}\delta_{ij}$$

取$\begin{cases}\delta_{ij}=1\\ \delta_{ji}=0\end{cases}$或者$\begin{cases}\delta_{ij}=0\\ \delta_{ji}=1\end{cases}$，不等式$d_j^{(t)}>d_i^{(t)}d_j^{(t)}$或者$d_i^{(t)}>d_i^{(t)}d_j^{(t)}$显然成立。证毕。

定理 2 表明,若两个机器人不能协调成功,则只需令推理置信度在 0 和 1 两个值上调整。这里,采用随机调整策略,即协调不成功按二值平均分布选取 0 或者 1,直到协调成功则保持推理置信度不变。这样,协调博弈问题由动作空间的搜索转化为参数空间的调整,缩减了搜索空间,提高了算法效率。

5　机器人避碰实验及结果分析

实验环境是 10×10 带障碍物的格子世界,环境中分布着 4 个机器智能体 $A_1,\cdots,A_4$,其动作集合为{不动,上移,右移,下移,左移},对应动作编号为 0,1,2,3,4。$A_1,\cdots,A_4$ 各自目标状态分别为 $G_1,\cdots,G_4$。机器智能体的目标是学习一条从起始点到目标状态的最短路径,同时避免和其余智能体碰撞。假定机器人只能观察到 4 个邻接方向上其余机器人的动作。

G_1	02	03	04	05	06	07	08	09	A_4
11	12	13	14	15	A_3	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	A_1	35	36	37	38	39	40
G_2	42	A_2	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	G_4	98	99	G_3

图 3　机器人所处的格子世界

每个智能体学习过程分两个阶段:在第一个阶段采用 Q 强化学习得到到达目标状态的最短路径(见表 3),在第二个阶段机器人共享第一阶段学习的 Q 表，采用 PEL 学习避碰行为。本文关注第二阶段的学习,并且在此阶段并不改变第一阶段的 Q 表值,机器人的学习目标是不发生碰撞的情况下,仍然按照第一阶段得到的最短路径行走，并且尽快到达各自目标。这里,假定机器人在任意位置等待的代价 c_1(比如耗电)小于其碰撞代价 c_2。在第二阶段,机器人动作集为{不动,前进},前进动作表示沿第一阶段最短路径行走一步。机器人的目标是协调相互行为，因此假定机器人之间存在利益共享(成本分担)机制。在每个碰撞点,博弈的支付矩阵如图 4 所示。

表 3　第一阶段学习到的最短路径

	最短路径
A_1	34–33–23–22–21–11–00;34–24–23–22–21–11–00
A_2	43–33–23–22–21–0–41
A_3	25–26–27⋯–77–78–79–80–90–100
A_4	10–20⋯–80–79–78–77–87–97

	前进	不动
前进	$(-c_2,-c_2)$	$(-c_1/2,-c_1/2)$
不动	$(-c_1/2,-c_1/2)$	$(-c_1,-c_1)$

图 4　机器人在碰撞点的支付矩阵

有两种机器人碰撞的情况：一是两机器人处于相邻位置,同时向对方位置移动;另一种是两机器人虽然不相邻,但同时向同一位置移动。作为惩罚,机器人每碰撞一次,回退到前一位置。比如:若 A_1 和 A_2 第一个时间步在位置 33 碰撞,则 A_1 和 A_2 回退到起始位置;若第一个时间步 A_1 不动,A_2 移动到位置 33,在第 2 个时间步 A_1 和 A_2 相向移动导致碰撞,则 A_1 不动,A_2 回退到起始位置。只有当碰撞发生时,机器人记录与之碰撞的机器人的行为，并基于这个行为和所共享的 Q 表推理对手下一时间步的行为。表 4 给出了可能的碰撞位置,以及学习结束后机器人在该位置的推理置信度。

表 4 中列出的只是多个可能的学习结果中的一个。比如在位置 23 的推理置信度还可能为 $\delta_{12}=1,\delta_{21}=0$。需要说明的是,处于相邻位置发生碰撞的两个机器人的推理置信度调整是相对于各自的位置而言的,比如位置(78,79)机器人 A_3、A_4 发生碰撞,则 $\delta_{34}=1$ 表示 A_3 在位置 78 处对于 A_4 的推理置信度。

表 4　可能的碰撞位置及机器人的参数调整情况

位置(状态)	碰撞后动作	推理置信度	协调结果
23	$a_1=2$	$\delta_{12}=0$	A_1:1–0–4–4–4–1–1
	$a_2=3$	$\delta_{21}=1$	A_2:1–1–4–4–3–3
33	$a_1=2$	$\delta_{12}=1$	A_1:4–1–4–4–1–1
	$a_2=3$	$\delta_{21}=0$	A_2:0–1–1–4–4–3–3
(78,79)	$a_3=4$	$\delta_{34}=1$	A_3:3–2–3–3–3–3–3–2–2
	$a_4=2$	$\delta_{43}=0$	A_4:3–3–3–3–3–3–3–0–0
(79,80)	$a_3=4$	$\delta_{34}=1$	A_3:3–2–3–3–3–3–3–2–2–2–3–3
	$a_4=1$	$\delta_{43}=0$	A_4:3–3–3–3–3–3–0–0–0–0–3–4–4–4–3–3
(77,78)	$a_3=1$	$\delta_{34}=0$	A_3:3–2–3–3–3–3–0–0–0–0–3–2–2–2–3–3
	$a_4=2$	$\delta_{43}=1$	A_4:3–3–3–3–3–3–3–4–4–4–3–3

以碰撞比率(实际碰撞次数与可能碰撞次数之比)考察算法的协调性能。图 5 列出了三种学习算法随时间变化曲线。可以看出,PEL 算法通过换位推理有效预测了对手行动，因此大幅度提高了协调性能。

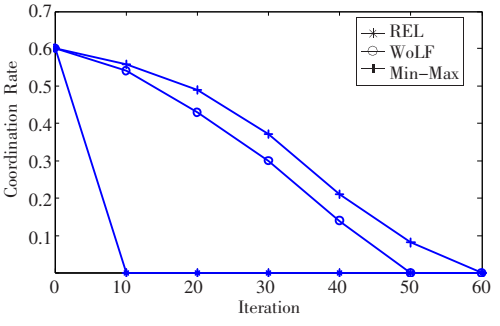


图 5　三种算法的协调性能对比情况

6　结束语

本文在分析协调博弈信念形成过程的基础上,利用智能体策略相似性,提出换位推理的协调博弈学习方法。智能体通过换位推理有效地预测了对手行为,通过信念修正模型将客观观察行为和主观预测行为结合在一起,从而能取得更好的协调性能。证明了初始信念不收敛情况下,通过调整推理置信度,智能体一定能协调成功。并且证明,智能体在一次协调成功之后,保

(下转 248 页)