ISSN1000-9825 Journal of Software 软件学报 2000,11(3):342~345

多 Agent 系统的几种规范生成机制*

王一川 石纯一

(清华大学计算机科学与技术系 北京 100084)

E-mail: wangyc@est4.cs.tsinghua.edu.cn

摘要 HCR(highest cumulative reward)是多 agent 系统中的一种规范生成机制,但在该机制下,系统的规范不能随条件的变化而变化.文章建立了规范的定义,分析了规范的稳定性,给出了用于规范生成的 HAR(highest average reward)和 HRR(highest recent reward)机制,适于规范的演化,并比 HCR 机制有更好的收敛速度.

关键词 多 agent 系统,协调,规范,突现行为,演化.

中图法分类号 TP18

行为规范是 agent 协调机制的一种.Agent 在交互时,根据行为规范在多个可能的行为之间直接作出选择,从而减少通信和协调开销.行为规范可以由设计者事先规定,也可以在某种机制的约束下由系统在运行中生成[1,2].后者又可分为两类,一类由中央控制结点来确定行为规范,另一类不存在中央控制的特权结点.各 agent 地位平等,在交互过程中逐渐生成规范,具有灵活、实现简单等优点.下面的分析只针对后一类.MAS 中的规范有生成、稳定和演化等过程,前者涉及行为策略的传播和行为的传播,后两者与系统达到规范状态后的变化有关.规范形成后,当某些 agent 为获得更高的短期收益而违反规范时,遵循规范的其他 agent 收益发生变化,此时,规范可能表现出保持稳定、产生波动或者解体的变化,这是规范的稳定过程.与之相似,当系统中未知行为被发现,使得系统中出现更优的行为策略时,规范向新行为策略的转化是规范的演化过程.规范的生成过程从局部来看,也是一个演化的过程.此外,采用规范的目的是在保证优化的同时减少协调开销,因而要求最终在全局范围内 agent 行为策略相同[3,4]是不必要的.

以往提出的规范生成机制可分为价值机制[3,5,6]和其他机制[7],两者都未能考虑到规范的稳定过程和演化过程.文献[3]在对策论的框架下定义了社会规范,用 HCR(highest cumulative reward)机制来保证在各种情形下生成有效的规范,agent 以历次交互中的各行为策略的累积收益为选择行为的依据.HCR 机制的缺点是不利于演化,同时,agent 必须预先知道行为策略集.文献[6]在 HCR 的基础上研究了 agent 交互的局部性和权威性对于规范生成结果的影响,以及在树型或层次型组织结构下规范的生成过程.这种方式仍有 HCR 机制的缺陷,并且组织结构中不同层次的 agent 计算能力由设计者预先设定,不能在动态过程中保证其合理性.文献[7]利用模仿机制研究了 agent 规范生成过程与收敛性、收敛速度有关的几个参数.Agent 记录每次交互时对方的行为策略,并根据历史信息来选择当前策略,其缺点是系统只会收敛到初始概率较高的策略.文献[5]研究了行为策略的传播过程,并比较了不同的传播方式的效率和可行性,但未涉及 agent 个体对行为策略的评价和取舍过程.

1 规范的定义

我们先给出几个必要的概念:

$E=(e_1,e_2,\ldots,e_n)$,其中 e_i 是给定的第 i 个 agent 交互时所处的场景,e_i 与 $e_j(i\neq j)$ 可能是相等的;C 为 E 中所有不同元素构成的集合;$A=\{a_1,a_2,\ldots,a_p\}$ 为交互中可选行为集;$F=\{f_1,f_2,\ldots,f_q\}$ 为行为策略集,其中

* 本文研究得到国家自然科学基金(No.69773026,69733020)资助.作者王一川,1973 年生,博士生,主要研究领域为分布式人工智能.石纯一,1935 年生,教授,博士生导师,主要研究领域为人工智能应用基础.

本文通讯联系人:王一川,北京 100084,清华大学计算机科学与技术系

本文 1998-12-01 收到原稿,1999-03-11 收到修改稿

$f_i:C\to A$.

效用函数 $u:A^n\to R^n$,$u(\alpha_{i1},\alpha_{i2},\ldots,\alpha_{in})=(u_{i1},u_{i2},\ldots,u_{in})$,其中 u_{ij} 为参与交互的 agent 在场景 e_j 下实施行为 α_{ij} 得到的收益;总效用函数 $U:A^n\to R$,$U(\alpha_{i1},\alpha_{i2},\ldots,\alpha_{in})=u_{i1}+u_{i2}+\ldots+u_{in}$.

F 上等价关系"$\approx$",$f_i\approx f_j\Leftrightarrow U(f_i(e_1),f_i(e_2),\ldots,f_i(e_n))=U(f_j(e_1),f_j(e_2),\ldots,f_j(e_n))$,由此得到 F 集上的一个划分 $\{F_1,F_2,\ldots,F_s\}$,不妨设其系统效用依次递增.定义 F 集上相容子集 Fc:对任意 $f_{i1},f_{i2},\ldots,f_{in}\in Fc$,有 $U(f_{i1}(e_1),f_{i2}(e_2),\ldots,f_{in}(e_n))\equiv U_{Fc}$,$U_{Fc}$ 为一常数.显然,相容集中各 f 等价.令 F 上所有相容集的集合为 FC.

定义规范:$Conv\in FC$,即规范是 F 上任一相容子集,相容性用以保证行为策略间不发生冲突.如果 $Conv$ 是最优等价集上的相容子集,则称该规范是全局优化规范.假定 agent 在交互中机会均等,则在一次遵循规范 $Conv$ 的交互前,agent 对自身收益的期望为 $eu=U_{Conv}/n$,此时全局优化的规范也满足 agent 的个体利益.

为了分析规范的稳定性,我们定义行为策略 f 对相容集 Fc 的优超:若存在 $k\in[1,n]$,$f_{j1},f_{j2},\ldots,f_{jn}\in Fc$,$f\in F$,有 $u(f_{j1}(e_1),f_{j2}(e_2),\ldots,f_{jn}(e_n))=(u_{j1},u_{j2},\ldots,u_{jn})$,$u(f_{j1}(e_1),f_{j2}(e_2),\ldots,f(e_k),\ldots,f_{jn}(e_n))=(u_{j1'},u_{j2'},\ldots,u_{jk'},\ldots,u_{jn'})$,且 $U(f_{j1}(e_1),f_{j2}(e_2),\ldots,f(e_k),\ldots,f_{jn}(e_n))<U_{Fc}$ 和 $u_{jk'}>u_{jk}$,即实施行为策略 f 的 agent 在损害整体收益的同时增加自身收益,此时称 f 优超相容集 Fc.若对于某相容集 F 不存在这样的 f,则 Fc 是稳定的;若存在某个 f,对任意 $f_{j1},f_{j2},\ldots,f_{jn}\in Fc$ 都有 $U(f_{j1}(e_1),f_{j2}(e_2),\ldots,f(e_k),\ldots,f_{jn}(e_n))<U_{Fc}$ 和 $u_{jk'}>u_{jk}$ 成立,则 Fc 是不稳定的;其他的情形介于稳定和不稳定之间,并可根据被优超的情况定义其稳定程度.显然,相容集在稳定性上优于其真子集.

2 HAR 和 HRR 算法

HCR 机制不能够满足演化的要求,因而在基于传播的规范生成过程中不能保证收敛.我们给出 HAR(highest average reward)和 HRR(highest recent reword)机制来消除累积效应,适合于规范的收敛和演化.HAR 以历史信息中各行为策略的平均收益作为选择当前行为策略的依据,用平均值来替代 HCR 中的累积值.HRR 在累计历史信息时利用归一化后的加权系数,给予越近发生的收益以越高的权值,由于归一化而消除了 HCR 的累积效应.但对于潜在规范不稳定的情形,HAR 和 HRR 也不能使系统收敛到规范.

下面给出算法 HAR 和 HRR.设 $C=\{e_1,e_2,\ldots,e_m\}$,$A=\{a_1,a_2,\ldots,a_q\}$,并假设 agent 初始时知道所有可行行为.

算法 1. HAR

定义收益数组 $reward[m][q]$,用来累积在某场景下采用某行为的收益;当前策略 $curStr[m]$,用于记录当前策略中 m 场景下所对应的行为;交互次数数组 $times[m][q]$,用于记录在某场景下采用某行为的次数.

(1) 初始化

$reward[m][q]$所有元素值设为某较大值,使得在各场景下不同的行为都有机会被执行.

$curStr[m]$随机初始化为$[1,q]$区间上的任意值.

$times[m][q]$各元素初始化为某正常数.其直观含义为 agent 对自身判断的信任程度.

(2) 每次交互,执行以下几步:

(a) 根据当前场景 e_i,得到当前策略下的对应行为 $a_{curStr[i]}$;

(b) 执行行为 $a_{curStr[i]}$,得到收益 cur_u;

(c) $reward[i][curStr[m]]$增加 cur_u;

$times[i][curStr[m]]$增 1;

(d) 如果 $cur\_u<reward[i][curStr[m]]/times[i][curStr[m]]$,则重新选择在场景 e_i 下应采取的行为 a_j,使得对任意 $k\in[1,q]$,有 $reward[i][j]/times[i][j]\geqslant reward[i][k]/times[i][k]$;令 $curStr[i]=j$.

算法 2. HRR

定义收益数组 $reward[m][q]$用于记录累积加权收益;当前策略 $curStr[m]$含义同算法 HAR;取定 $weight\in[0,1]$为加权比.

(1) 初始化

$reward[m][q]$所有元素值设为某较大值,使得在各场景下不同的行为都有机会被执行.

$curStr[m]$随机初始化为$[1,q]$区间上的任意值.

(2) 每次交互,执行以下几步:

(a) 根据当前场景e_i,得到当前策略下的对应行为$a_{curStr[i]}$;

(b) 执行行为$a_{curStr[i]}$,得到收益cur_u;

(c)* $reward[i][curStr[m]]=reward[i][curStr[m]]\times weight-cur\_u\times(1-weight)$;

(d) 如果$cur_u<reward[i][curStr[m]]$,则重新选择在场景$e_i$下应采取的行为$a_j$,使得任意$k\in[1,q]$,有$reward[i][j]\geq reward[i][k]$;令$curStr[i]=j$.

3 实验分析

3.1 实验说明

实验背景由100个agent组成,agent两两随机交互,用于检验行为规范生成机制的收敛性和演化性.收敛的标准是连续1 000次交互中按规范进行的次数大于950次,并以其最早出现的时刻为收敛时刻.每个实验限定交互次数为8 000次,各做1 000回.

实验1. 正值收益情形下各机制的收敛性

$E=(e_1,e_1);A=\{a_1,a_2\};u(a_1,a_1)=u(a_2,a_2)=(4,4),u(a_2,a_1)=u(a_1,a_2)=(1,1);F=\{f_1,f_2\},f_1(e_1)=a_1,f_2(e_1)=a_2$.

初始时每个agent知道全部两种行为.

HRR中取$weight=0.8$**.

实验2. 不同初始概率下各机制的收敛性

$E=(e_1,e_1);A=\{a_1,a_2\};u(a_1,a_1)=(4,4),u(a_2,a_2)=(1,1),u(a_2,a_1)=u(a_1,a_2)=(-1,-1);F=\{f_1,f_2\},f_1(e_1)=a_1,f_2(e_1)=a_2$.

初始时每个agent只知道一种策略,该策略是f_1的概率,为P,agent在交互中相互传播策略.当P较小时,系统向f_1的收敛过程也是一个规范演化的过程.

HRR中取$weight=0.8$.HCR中已知而未尝试的行为策略在选择时有优先权.

3.2 实验结果

在实验1中,HAR和HRR算法在1 000回中全部收敛,而HCR算法均不能收敛.

实验2的结果见表1,表中内容为在标定交互次数内1 000回测试所收敛的回数.可以看出,当P不同时,HCR的收敛性的变化很大,P越大,收敛性越好.HAR和HRR算法对不同的P都有很好的收敛性.

Table 1

表1

	1 000	2 000	3 000	4 000	5 000	6 000	7 000	8 000
HCR $P=1/2$	997	999	999	1 000	1 000	1 000	1 000	1 000
HCR $P=1/3$	650	778	811	820	830	831	834	836
HCR $P=1/4$	178	284	328	345	357	363	365	367
HCR $P=1/6$	9	29	35	38	38	40	40	40
HCR $P=1/25$	0	0	0	0	0	0	0	0
HAR $P=1/2$	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000
HAR $P=1/25$	954	976	976	976	976	976	976	976
HRR $P=1/2$	1 000	1 000	1 000	1 000	1 000	1 000	1 000	1 000
HRR $P=1/25$	977	977	977	977	977	977	977	977

* HRR中加权系数总和为1,近远相邻两次交互收益的加权系数比为$1/weight$.这里,通过引入agent对行为的初始收益而对算法作了适当的简化.

** $weight$值越小,当前值的比重越大,系统振荡的机会也越大.0.8是一个经验值.

3.3 结果分析

当 agent 知道效用函数所确立的映射时，可以将整个收益的计算平移到零值左右，此时，HCR 机制仍可能起作用. 但在实际情况中，agent 很难预先知道由系统决定的效用函数，实验 1 的意义就在于指出 HCR 机制对 agent认知能力的这种要求，而 HAR 和 HRR 机制没有这个限制.

实验 2 中，HCR 机制的收敛性与 P 值有关，P 越大，收敛性越好；P 值较小时，初始策略为 f_2 的 agent 能够在互相交互中积累足够的收益来阻碍 f_1 的再次被选取，同时也使当前策略为 f_1 的 agent 互相交互的可能性减少.

4 结 语

本文以无冲突为基点建立了多场景下规范的一般模型，通过分析规范的生成过程，给出了两种基于价值选择的规范生成机制，并通过分析实验比较了几种机制在简单情形下的收敛性. 本文没有讨论 agent 对于场景的识别而假定 agent 内在具备识别场景的能力，但在一些情形下，场景的识别与行为的收益有关，和规范的生成互相影响，使得规范的生成过程更加复杂. 实验中不存在优超行为策略，因而没有检验规范生成机制的稳定性.

参考文献

1 Ephrati E, Pollack M E, Ur S. Deriving multi-agent coordination through filtering strategies. In: Mellish C S ed. Proceedings of the 14th International Joint Conference on Artificial Intelligence, Vol 1. San Mateo, CA: Morgan Kaufmann Publishers, 1995. 679～685

2 Goldman C V, Rosenschein J S. Emergent coordination through the use of cooperative state-changing rules. In: Proceedings of the 12th National Conference on Artificial Intelligence, Vol 1. Cambridge, MA: MIT Press, 1994. 408～413

3 Shoham Y, Tennenholtz M. On the emergence of social conventions: modeling, analysis, and simulations. Artificial Intelligence, 1997,94(1～2):139～166

4 Tennenholtz M. On stable social laws and qualitative equilibria. Artificial Intelligence, 1998,102(1):1～20

5 Luo Yi. Agent model and solving method in multi-agent system [Ph.D. Thesis]. Beijing: Tsinghua University, 1996
(罗颖. 多 Agent 系统中 Agent 模型和求解方法[博士学位论文]. 北京：清华大学，1996)

6 Kittock J E. The impact of locality and authority on emergent conventions: initial observations. In: Proceedings of the 12th National Conference on Artificial Intelligence, Vol 1. Cambridge, MA: MIT Press, 1994. 420～425

7 Walker A, Wooldridge M. Understanding the emergence of conventions in multi-agent systems. In: Lesser V, Gasser L. eds. Proceedings of the 1st International Conference on Multi-Agent Systems. Cambridge, MA: MIT Press, 1995. 384～389

Strategy-Selection Rules for Developing Conventions in Multi-Agent System

WANG Yi-chuan SHI Chun-yi

(*Department of Computer Science and Technology Tsinghua University Beijing* 100084)

Abstract Highest cumulative reward (HCR) is a rule for developing conventions in multi-agent systems. But it will keep system maintaining an emerged convention from evolving to more rational ones while conditions of system are developing. In this paper, the notion of conventions is defined, and the stability of them is analyzed. Furthermore, two rules called highest average reward (HAR) and highest recent reward (HRR) are introduced. They both guarantee the evolving process of stable conventions, and the convergence rate of them is better than that of HCR.

Key words Multi-Agent system, coordination, convention, emergent behavior, evolution.