第 25 卷第 5 期
2008 年 10 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 25 No. 5
Oct. 2008

文章编号: 1000−8152(2008)05−0853−04

多任务联盟形成中的Agent行为策略研究

蒋建国[1,2], 苏兆品[1,2], 张国富[1,2], 夏 娜[1,2]

(1. 合肥工业大学 计算机与信息学院, 安徽 合肥 230009; 2. 安全关键工业测控技术教育部工程研究中心, 安徽 合肥 230009)

摘要: Agent联盟是多Agent系统中一种重要的合作方式, 联盟形成是其研究的关键问题. 本文提出一种串行多任务联盟形成中的Agent行为策略, 首先论证了Agent合作求解多任务的过程是一个Markov决策过程, 然后基于Q-学习求解单个Agent的最优行为策略. 实例表明该策略在面向多任务的领域中可以快速、有效地串行形成多个任务求解联盟.

关键词: 串行多任务; 联盟; Agent行为策略; Q-学习

中图分类号: TP301 **文献标识码**: A

Agent-behavior strategy in serial multi-task coalition formation

JIANG Jian-guo[1,2], SU Zhao-pin[1,2], ZHANG Guo-fu[1,2], XIA Na[1,2]

(1. School of Computer and Information Science, Hefei University of Technology, Hefei Anhui 230009, China;
2. Engineering Research Center of Safety Critical Industrial Measurement and Control Technology,
Ministry of Education, Hefei Anhui 230009, China)

Abstract: Agent-coalition is an important approach to agent-coordination and cooperation, in which the coalition formation is a key topic. Existing researches are restricted in single-task environments, and the results are not applied to multi-task environments. In this paper, a new agent behavior strategy in serial multi-task coalition formation for problem-solving is presented. The conclusion shows that the agent-task selection is a Markov Decision Process. The Q-learning is used to optimize the behavior strategy for a single agent, and the cooperative multi-agent reinforcement learning improves the learning rate. Experiments prove that the strategy can effectively and serially form coalitions for multi-task.

Key words: serial multi-task; coalitions; Agent behavior strategy; Q-learning

1 引言(Introduction)

在多Agent系统(multi-Agent systems, MAS)中, 由于Agent的自身知识和能力有限, 需要组成联盟求解复杂任务而获得效用.

自1993年提出“联盟”概念以来, 联盟形成已成为MAS研究的一个重要方面. 近年来的研究主要是如何在联盟内Agent间划分联盟的收益. 已有多数收益划分方案是基于N人合作对策论, 如核、Shapley值等. Shehory等人[1]的工作具有代表性. 但这些研究没有考虑算法和计算资源、计算分布等要求, 只考虑解的存在性, 而且不保证全局最优和联盟稳定. 罗翊[2]提出了非减性收益分配原则, 具有简单时效等优点, 但平均分配没有反映出Agent贡献的差异性, 且契约法则打击了其他Agent加入已有联盟的兴趣. 文献[3]基于拍卖和合同机制的形成策略, 虽然既严格遵循按劳分配又完全符合效用非减, 可以形成全局最优联盟, 但只限于考虑单一Agent联盟和单一任务.

上述研究都没有考虑在联盟形成过程中的Agent自身行为. 文献[4]基于联盟的统计规律提出Agent熟人的概念, 并提出了以熟人为基础的联盟策略和联盟竞争任务的算法, 该策略在一定程度上减少了通讯开销和计算量, 但随着熟人集的规模减小将会漏掉较优解, 且在搜索的初期计算量也比较大. 文献[5]把Agent选择子任务的过程看成马尔科夫决策过程, 但此方法只适用于Agent不能完全控制其能力消耗的不确定环境下, 而且获得最优策略的速度较慢. 本文在已有研究的基础上, 针对串行多任务联盟形成问题提出一种基于Q-学习的Agent行为策略, 并通过实例说明该策略的有效性.

收稿日期: 2007−03−23; 收修改稿日期: 2007−12−25.

基金项目: 国家自然科学基金资助项目(60474035); 国家教育部博士点基金资助项目(20060359004); 安徽省自然科学基金资助项目(070412035).

2 问题描述(Problem description)

MAS中存在m个理性的Agent, 表示为$Agent = \{1, 2, \cdots, m\}$, 任意Agent $k \in Agent$都有一个能力向量: $B_k = (b_k^1, b_k^2, \cdots, b_k^r)$ $(b_k^j \geqslant 0,\ j = 1, 2, \cdots, r)$, 用于定量描述$k$执行某种特定动作的能力大小. 现有$N$个依次待求解的任务$T = \{T^1, T^2, \cdots, T^N\}$, 每个任务$T^i$ $(i = 1, 2, \cdots, N)$具有一定的能力需求$B_{T^i} = (b_{T^i}^1, b_{T^i}^2, \cdots, b_{T^i}^r)$, $(b_{T^i}^j \geqslant 0)$. Agent完成任务T^i可获得相应的利益$P(T^i)$, Agent k 参与任务T^i求解的必要条件为$\exists j \in \{1, 2, \cdots, r\}$, 使得$b_k^j \geqslant b_{T^i}^j$.

这里的理性[6,7]是指Agent在面对多个可能的行为策略时会根据自己的分析作出合理的选择. 一般认为, 理性Agent具有理性化的思维能力, 但同时也是贪婪自私、唯利是图的, 以最大化自身收益为目标. 因而理性Agent间更倾向于组成联盟, 从而提高求解效率, 获得更多的报酬.

任务T^i的求解联盟$\mathbf{C}_{T^i}$用$\langle C_{T^i}, \mathrm{alloc}_{T^i}, u_{T^i}\rangle$三元组表示, 其中$C_{T^i} \subseteq Agent$且$C_{T^i} \neq \phi$, 联盟$C_{T^i}$的收益用一个特征函数$V(C_{T^i})$给出, 如公式(1)所示, 其中$\varphi(T^i, k, j)$满足公式(2).

$$V(C_{T^i}) = P(T^i) - \sum_k \sum_j b_k^j \cdot \varphi(T^i, k, j), \tag{1}$$

$$\varphi(T^i, k, j) = \begin{cases} 1, \text{Agent } k\text{参与完成}T^i\text{的第}j\text{维能力}, \\ 0, \text{否则}. \end{cases} \tag{2}$$

联盟C_{T^i}中的成员共同分享收益$V(C_{T^i})$, 收益分配向量用$u_{T^i} = (u_{T^i}^1, u_{T^i}^2, \cdots, u_{T^i}^{|C_{T^i}|})$表示, $u_{T^i}^k$为Agent k所获得的收益(如公式(3)和(4)), 且满足$\sum\limits_{k \in C_{T^i}} u_{T^i}^k = V(C_{T^i})$, $g_k(b_{T^i}^j)$表示Agent k完成任务T^i的第j维能力所获得的收益.

$$u_{T^i}^k = \sum_j g_k(b_{T^i}^j)\varphi(T^i, k, j), \tag{3}$$

$$g_k(b_{T^i}^j) = \frac{b_{T^i}^j}{\sum\limits_{l=1}^{r} b_{T^i}^l} V(C_{T^i}). \tag{4}$$

alloc_{T^i}为任务分配函数, 若 $b_k^j \geqslant b_{T^i}^j$, 则$\mathrm{alloc}_{T^i}(b_{T^i}^j) = k$. 联盟$C_{T^i}$能够承担任务$T^i$的必要条件是对于任意一维能力需求$b_{T^i}^j$, 都必然至少存在一个Agent $k \in C_{T^i}$, 满足$b_k^j \geqslant b_{T^i}^j$.

为了获得联盟的最大收益, 引入系统控制者来确定最终的任务求解联盟$C_{T^i}^*$, 即

$$V(C_{T^i}^*) = \max V(C_{T^i}). \tag{5}$$

串行多任务联盟形成问题就是为任务集$T = \{T^1, T^2, \cdots, T^N\}$ 中的任务依次形成最优或次优求解联盟, 并能使得联盟中各Agent的个体收益最大.

3 串行多任务联盟形成中的Agent行为策略(Agent behavior strategy in serial multi-task coalition formation)

在多任务环境中, Agent参与求解的任务越多, 获得的收益也就越大; 但Agent的能力是有限的, 不可能参与所有任务的求解. 因此, 自私的Agent为了最大化自身的收益, 只能理性地选择参与多个任务求解联盟. 从本质上来讲, 串行多任务联盟形成的过程就是各Agent自主选择任务进行求解的过程.

3.1 Agent的任务选择过程(Process of Agent task selection)

对于待求解的任务$T^i \in T$, Agent $k \in Agent$根据现有的能力面临两种选择: 加入联盟获得收益; 或者节省能力, 参与后来任务的求解. 选择的结果会使Agent的能力和已参与的任务发生变化.

定义 1 Agent的状态[5]: 用$\langle RA, TA\rangle$二元组表示Agent所处的状态, $RA = (ra^1, ra^2, \cdots, ra^r)$为Agent当前所能提供的能力向量, 即剩余能力向量; TA表示Agent已参与求解的任务集合. Agent 所有可能处于的状态s组成的集合称为状态集, 表示为$S = \{s_i | i = 1, 2, \cdots\}$.

由于Agent能完全控制自身能力的消耗, 对于Agent$k$, 当面对任务T^i时, 根据自身能力与任务能力需求, 只可能进行两种操作: ① 当$ra_k^j \geqslant b_{T^i}^j$时, 参与$T^i$的求解, 消耗一定的能力, 获得收益, 用$J$表示; ② 当$ra_k^j < b_{T^i}^j$时, 不参与$T^i$的求解, 保存能力, 参与下一任务的求解, 用$K$表示. 因此

定义 2 Agent的动作集合$Action = \{J, K\}$.

Agent在状态s_i执行动作$a \in Action$, 转移到下一个状态s_{i+1}, s_{i+1}与s_i 的关系可以表示为(6)(7)和(8).

$$s_{i+1} = \langle RA_{i+1}, TA_{i+1}\rangle, \tag{6}$$

$$RA_{i+1} = \begin{cases} RA_i - \sum\limits_j (0, \cdots, b_{T^i}^j, \cdots, 0), & a = J, \\ RA_i, & a = K, \end{cases} \tag{7}$$

$$TA_{i+1} = \begin{cases} TA_i \bigcup \sum\limits_j \{b_{T^i}^j\}, & a = J, \\ TA_i, & a = K. \end{cases} \tag{8}$$

由于Agent对每一个任务都要进行选择, 因此Agent选择任务求解的过程是一个顺序决策过程. 可以得出:

定理 1 由状态$s_i = \langle RA_i, TA_i\rangle$ $(i \geqslant 0)$所构成的决策过程是一个马尔科夫决策过程, 即Agent选择任务求解的过程是一个马尔科夫决策过程.

证　由马尔科夫决策过程[8]的定义可知, 只要证明状态s_{i+1}仅仅依赖于s_i. 由公式(6)(7)和(8)可知, Agent$k$在$i+1$时刻的剩余能力RA_{i+1}仅由i时刻的剩余能力RA_i决定, 且TA_{i+1}也依赖于TA_i. 因此, 状态s_{i+1}仅由s_i决定, 与之前的其他任何状态无关, Agent选择任务求解的过程是一个马尔科夫决策过程.　　证毕.

3.2　基于Q-学习的Agent 最优行为策略(Optimal Agent behavior strategy based on Q-learning)

Q-学习[9]是由Watkins提出的一种模型无关的强化学习算法, 设环境是一个有限状态的离散马尔科夫决策过程, 每步可在有限动作集合中选取某一动作, 环境接受该动作后状态发生转移, 同时给出评价. Q-学习实际上是马尔科夫决策过程的一种变化形式.

定义 3　Agent的状态值$v(s)$: 是指Agent在达到s以前参与求解任务所得到的收益之和.

定义 4　Agent的瞬时奖赏$r(s')$: 指Agent从s转移到s', 执行动作a所获得的收益. 例如Agent k参加了任务T^{i+1}的第j维能力的求解, 获得的收益为$g_k(b^j_{T^{i+1}})$, 则$r(s')=g_k(b^j_{T^{i+1}})$.

由定理1, Agent选择任务求解的过程是一个马尔科夫决策过程. 在此过程中, 各Agent的目标是在每个离散状态$s \in S$, 即每个Agent面临的任务, 选择最优行为策略$a \in Action$, 以使自身的期望收益值最大. 在行为策略π的作用下, 状态s_i的值如式(9), 其中γ为折扣因子. 动态规划理论保证至少有一个策略π^*, 使得式(10)成立. 学习的任务就是确定π^*, 使总的收益期望值最大.

$$v^{\pi}(s)=r(\pi(s))+\gamma\sum_{s'\in S}P(s,a,s')v^{\pi}(s'), \quad (9)$$

$$v^{\pi^*}(s)=\max_{a\in Action}\{r(\pi(s))+\gamma\sum_{s'\in S}P(s,a,s')v^{\pi^*}(s')\}. \quad (10)$$

其中$P(s,a,s')$为系统在状态s执行动作a后使系统状态转移到s'的概率, 简称为状态转移概率. Q-学习的思想是不去估计环境模型, 而是直接优化一个可迭代的Q函数. Q函数是在s 时执行a, 且以后一直遵循最优动作序列, 其调整方式如为式(11), 其中α为学习速度. 算法的收敛性已经得到了证明[9], 使其具有了更可靠的理论支持. 但因为Agent在每次学习迭代时都需要考察每一个行为, 使得收敛速度较慢. 为了提高收敛速度, 引入合作多Agent强化学习(cooperative multi-Agent reinforcement learning)方法. 该方法是由Tan提出的一种多Agent强化学习方法, 通过交换Agent的状态信息、学习经验和策略等手段快速增加各Agent的知识, 提高学习的速度[10].

$$Q_{i+1}(s,a)\leftarrow (1-\alpha)Q_i(s,a)+\alpha[r(s)+\gamma\max_{a'\in Action}Q_i(s',a')]. \quad (11)$$

Agent为了能以较大概率被系统控制者选择加入最终求解联盟, 获得尽可能大的收益, 需要与其他Agent相互交换状态信息, 了解其他Agent的剩余能力, 再确定自身的状态转移概率. 当Agent面向任务T^i时, 由于Agent的自私性, 每个Agent均不想被其他Agent明确感知到自己剩余能力的多少, 所以只会以“能”或“不能”完成任务的方式与其他Agent交互. 因此引入持续向量的定义.

定义 5　Agent$k$的持续向量(Duration): $Dur_k=(d^1_k,d^2_k,\cdots,d^r_k)$, d^j_k的取值由式(12)确定. 每求解一个新的任务, 更新Dur_k.

$$d^j_k=\begin{cases}1, \text{ 若 } ra^j_k \geqslant b^j_{T^i},\\ 0, \text{ 否则.}\end{cases} \quad (12)$$

Agent在学习过程中与其他Agent不断交互持续向量的值, 了解彼此的状态信息, 以决定自身的状态转移. 因此Agent k获得最优行为策略的步骤为:

Step 1　确定当前的状态s_i;

Step 2　选择动作a, 并执行a;

Step 3　确定下一个状态s_{i+1};

Step 4　获得即时收益r;

Step 5　按照公式(11)更新Q值;

Step 6　令$s_i=s_{i+1}$, 转向Step2, 直到学习结束, 得到最优的策略.

3.3　联盟形成过程(Coalition formation process)

串行多任务联盟的形成过程可以描述为:

Step 1　每个Agent通过Q学习确定其要求解的任务和相应的能力贡献向量;

Step 2　每个Agent将最终状态提交给系统, 系统控制者依次确定每个任务的求解联盟;

Step 3　根据公式(3)和(4)确定Agent的收益.

对于串行多任务联盟的形成问题, 要想找出全局最优解, 最根本的是由系统搜索所有可能的联盟, 费时费力, 不能满足实际系统的需求. 通过引入Q-学习, 先由Agent确定自己的最优行为策略, 然后再提交给系统, 这样可大大减少系统全局搜索的时间和空间, 因此本文的方法可以找到系统的全局最优解.

4　实例分析(Example)

设有4个待求解的任务$T=\{T^1,T^2,T^3,T^4\}$, 每个任务的能力需求分别为:$B_{T^1}=(3,5,8)$, $B_{T^2}=(6,10,3)$, $B_{T^3}=(2,4,5)$和$B_{T^4}=(6,9,6)$, 完成4个任务所能获得的利益分别为:$P(T^1)=32$,

$P(T^2)=38$, $P(T^3)=22$和$P(T^4)=42$; MAS中存在8个理性Agent, $Agent=\{1,2,\cdots,8\}$, 每个Agent k的能力向量为B_k(由于篇幅限制, 不予列出). 按照本文的策略形成多任务求解联盟的步骤如下:

1) 根据3.2, 基于Q-学习确定每个Agent期望参与的任务和相应的能力贡献向量, 如表1.

表 1 Agent期望参与的任务和相应的能力贡献向量

Table 1 Tasks that Agents expect to join and the corresponding ability vector

Agent	任务	能力贡献量
1	$b^3_{T^2}$	$(0,0,4)$
2	$b^1_{T^1}, b^2_{T^1}, b^3_{T^3}$	$(3,0,0), (0,6,0), (0,0,5)$
3	$b^1_{T^2}b^1_{T^3}, b^2_{T^2}, b^3_{T^4}$	$(6,0,0)(2,0,0), (0,11,0), (0,0,7)$
4	$b^1_{T^2}, b^2_{T^1}, b^3_{T^4}$	$(7,0,0), (0,7,0), (0,0,6)$
5	$b^1_{T^1}, b^2_{T^1}, b^3_{T^2}b^3_{T^4}$	$(4,0,0), (0,6,0), (0,0,3)(0,0,6)$
6	$b^1_{T^1}b^1_{T^2}, b^2_{T^1}b^2_{T^3}, b^3_{T^4}$	$(3,0,0)(6,0,0), (0,5,0)(0,4,0), (0,0,7)$
7	$b^1_{T^2}b^1_{T^3}, b^2_{T^3}b^2_{T^4}, b^3_{T^1}b^3_{T^2}$	$(6,0,0)(2,0,0), (0,4,0)(0,9,0), (0,0,8)(0,0,3)$
8	$b^1_{T^4}, b^2_{T^2}, b^3_{T^1}b^3_{T^2}$	$(7,0,0), (0,10,0), (0,0,8)(0,0,4)$

2) 系统控制者按照公式(5)依次为每个任务确定最终求解联盟及其联盟收益, 如表2.

表 2 任务的最终求解联盟及收益

Table 2 Coalitions for tasks and rewards

	任务	T^1	T^2	T^3	T^4
本文策略	联盟	$\{2,6,8\}$	$\{7,8\}$	$\{7,6,2\}$	$\{8,7,4\}$
	收益	13	18	11	18
文献[4]策略	联盟	$\{1,2\}$	$\{3\}$	$\{5\}$	$\{6\}$
	收益	12	12	3	17

3) 确定参加任务求解的Agent的收益, 如表3.

表 3 Agent的个体收益

Table 2 Utility for Agents in coalitions

Agent	2	4	6	7	8
收益	7.44	5.14	8.06	7.68	24.50

如表2, 采用文献[4]的策略为每个任务形成联盟的收益均小于本文的方法, 这是因为采用熟人集虽然能减少计算量, 但很可能漏掉较优解. 因此本文策略可以有效、快速地为串行多任务形成最优求解联盟, 在最大化联盟收益的同时, 充分考虑到了各Agent的自身行为, 满足了各Agent追求自身收益最大化的需求, 采用的收益分配方案符合按劳分配的思想, 在一定程度上确保了联盟的稳定性和求解任务的效率.

5 结论(Conclusion)

本文针对串行多任务环境中的联盟形成问题给出了一种新的Agent行为策略, 首先论证了Agent合作求解任务的过程是一个Markov决策过程, 基于Q-学习设计了单个Agent的行为策略, 并引入了合作多Agent强化学习方法来提高学习的速度. 实例说明该策略可以快速、有效地为串行多任务形成最优求解联盟, 在一定程度上满足了实际系统的需求.

参考文献(References):

[1] SHEHORY O, KRAUS S. Formation of overlapping coalitions for precedence-ordered task-execution among autonomous Agents[C] // *Proceedings of the 2nd International Conference on Multi-Agent Systems*. Kyoto, Japan: MIT Press, 1996: 330 – 337.

[2] 罗翊, 石纯一. Agent协作求解中形成联盟的行为策略[J]. 计算机学报, 1997, 20(11): 961 – 965.
(LUO Yi, SHI Chunyi. The behavior strategy to form coalition in Agent cooperative problem-solving[J]. *Chinese Journal of Computer*, 1997, 20(11): 961 – 965.)

[3] JIANG J G, ZHANG G F, SU Z P. Strategy of coalition formation based on distribution according to work and non-reducing utility[J]. *Chinese Journal of Electronics*, 2007, 16 (4): 573 – 577.

[4] 叶东海, 蓝少华, 王玉善, 等. 基于熟人的Agent联盟策略[J]. 小型微型计算机系统, 2000, 21(10): 1053 – 1055.
(YE Donghai, LAN Shaohua, WANG Yushan, et al. Strategy of Agent coalition based on acquaintance[J]. *Mini-Micro Systems*, 2000, 21(10): 1053 – 1055.)

[5] HOSAM H, KHALDOUNN Z. Planning coalition formation under uncertainty: auction approach[J]. *Information and Communication Technologies*, 2006, 2: 3013 – 3017.

[6] KLUSCH M, SHEHORY O. Coalition formation among rational information Agents[C] // *Proceedings of European Workshop Modeling Autonomous Agents in a Multi-Agent World*. Heidelberg, Germany: Springer-Verlag, 1996: 204 – 217.

[7] KLUSCH M, GERBER A. Dynamic coalition formation among rational Agents[J]. *IEEE Journal on Intelligent Systems*, 2002, 17(3): 42 – 47.

[8] BELLMAN R E. A markov decision process[J]. *Journal of Mathematical*, 1957, 6: 679 – 684.

[9] WATKINS P. Q-Learning[J]. *Machine Learning*, 1992, 8(3): 279 – 292.

[10] TAN M. Multi-Agent reinforcement learning: independent vs. cooperative Agents[C] // *Proceedings of the 10th International Conference on Machine Learning*. Alberst, MA, USA: Morgan Kaufmann Publisers, 1993: 330 – 337.

作者简介:

蒋建国 (1955—), 男, 教授, 博士生导师, 从事分布式智能控制系统、数字图像分析与处理、DSP技术与应用的研究, E-mail: szhpin@163.com;

苏兆品 (1983—), 女, 博士研究生, 从事智能控制、进化计算、Agent理论的研究;

张国富 (1979—), 男, 博士, 讲师, 从事分布式人工智能、不确定系统、进化计算的研究;

夏 娜 (1979—), 男, 副教授, 硕士生导师, 从事分布式人工智能、无线传感器网络的研究.