基于 Q-learning 的一种多 Agent 系统结构模型*

许 培 薛 伟
(江南大学物联网工程学院 无锡 214122)

摘 要 多 Agent 系统是近年来比较热门的一个研究领域,而 Q-learning 算法是强化学习算法中比较著名的算法,也是应用最广泛的一种强化学习算法。以单 Agent 强化学习 Q-learning 算法为基础,提出了一种新的学习协作算法,并根据此算法提出了一种新的多 Agent 系统体系结构模型,该结构的最大特点是提出了知识共享机制、团队结构思想和引入了服务商概念,最后通过仿真实验说明了该结构体系的优越性。

关键词 多 Agent 系统;强化学习;Q 学习;体系结构;知识共享
中图分类号 TP18

A Structure Model of Multi-agent System Based on Q-learning

Xu Pei Xue Wei
(Dept. of Internet of Things, Jiangnan University, Wuxi 214122)

Abstract Multi-agent system(MAS) is a very popular research field in recent years, and Q-learning algorithm is a more famous algorithm, also is one of the most widely used reinforcement learning algorithms. In this paper, a new algorithm based on learning cooperation is proposed. Its basis is Q-learning, a single agent reinforcement learning algorithm. Finally, a novel structure model of MAS is proposed. The most significant characteristic of the model is to put forward the mechanism of knowledge sharing, the ideal of team structure and to introduce the concept of facilitator. In the end, it shows the advantage of the structure system by the simulation experiment.

Key Words multi-agent system(MAS), reinforcement learning, Q-learning, structure system, knowledge sharing
Class Number TP18

1 引言

MAS 系统[1]是由多个智能 Agent 通过相互协调、相互作用、相互联系构成的系统,MAS 的重点在于使功能独立的 Agent 通过协商、协调和协作,完成复杂的控制任务或解决复杂的问题。目前,在多 Agent 系统理论中,多 Agent 之间的结构体系及学习协作是核心问题。多 Agent 系统的组织与控制方式对系统性能的影响极大,如何组织多 Agent 系统以及如何在系统中实现多 Agent 之间的学习协作问题,已成为当前多 Agent 系统领域的一个新课题,具有重要的理论和实际意义。

本文以 Q-learning 算法理论为基础,结合多 Agent 系统的特点及学习协作方式,提出了一种新的学习协作算法,并提出一种新的体系结构,最后通过仿真分析,来说明该结构体系的优越性。

2 强化学习 Q-learning 算法

强化学习是指从环境状态到行为映射的学习,以使系统行为从环境中获得的累积奖赏值最大,是一种以环境反馈作为输入的、特殊的、适应环境的机器学习方法。传统的强化学习中单 Agent 学习问题可以通过马尔可夫决策过程(Markov decision process,MDP)建模[2]。一个单 Agent MDP 被定义为一个四元组(S,A,P,R),其中 S 是一个有限的状态集,A 是一个有限的动作集,P 是环境的状

* **收稿日期**:2011 年 1 月 30 日,**修回日期**:2011 年 3 月 1 日
作者简介:许培,男,硕士研究生,研究方向:信息处理系统、智能系统。薛伟,男,副教授,硕士生导师,研究方向:嵌入式系统应用与智能控制。

态转移函数，R 是环境的奖励函数。记 $R(s_{t+1}|s_t, a)$ 为 Agent 在状态 s_t 采用 a 动作使环境状态转移到 s_{t+1} 获得的瞬时奖赏值；记 $P(s_{t+1}|s_t, a)$ 为 Agent 在状态 s_t 采用 a 动作使环境状态转移到 s_{t+1} 的概率。对于智能体的强化学习，其目的是学习一个行为策略 $P: S—A$，使其选择的动作可以获得最大的累积奖励 R。如果系统某个动作导致环境正的奖赏，那么系统以后产生这个动作的趋势便会加强，反之则减弱。

Q-learning 算法是强化学习方法中的一个比较著名的算法，也是强化学习中应用最广的一种算法，Q 函数的定义为在状态 s 时执行动作 a 且此后按最优动作序列执行时的折扣累计强化值[3]。智能体最优策略为在每一状态选用 Q 值最大的行为。$Q^*(s,a)$ 和最优策略 $\pi^*(s,a)$ 求解分别如式(1)和式(2)所示：

$$Q^*(s,a)=R(s,a)+\gamma\sum_{s'\in S}P(s'|s,a)\max_{a'\in A}Q^*(s',a') \tag{1}$$

$$\pi^*(s,a)=\arg\max_{a\in A,s\in S}Q^*(s,a) \tag{2}$$

Q^* 和 π^* 的定义遵循 Bellman 的最优化原理[4]，如果知道了 $Q^*(s,a)$，那么最优策略 π^* 就会被找到。目前 Q 学习法是最好的通过计算 Q 值，产生概率，并根据概率选择行动的单 Agent 强化学习方法。

3　多 Agent 系统学习协作算法

各 Agent 直接采用单 Agent 强化学习方法并通过某种协作机制实现协作是实现多 Agent 协作学习的一种思路。本文基于该思想下提出一种改进的多 Agent 协作学习方法。具体来说，就是各个 Agent 都采用单 Agent 强化学习算法进行独立学习，同时多 Agent 系统通过知识共享机制实现知识共享和协作操作，从而能够充分利用系统知识资源，起到加速学习和提高整个系统效率的作用。

3.1　知识共享机制

由于单 Agent 强化学习不考虑其他 Agent 的状态变化，故它的学习带有一定的自私性[5]，从而整个系统无法充分的协调配合工作，对于这个问题的出现，本文提出了知识共享机制，类似于高级语言中全局量结构体的概念。本文中单 Agent 强化学习采用的是 Q-learning 算法，每个 Agent 学习完后会得到一个对应于该状态下的一个值 $Q_t(s_t, a_t)$，于是，我们可以建立一个 Q 值表来存储得到的各个 Agent 的 Q 值，这里知识的一种直接的表达是 Q 表，Q 表说明了各个 Agent 在某种状态下行为的倾向。通过共享该 Q 表，来达到各 Agent 之间相交互的目的。知识共享结构如图 1 所示，当 Agent 在一个周期(t)的学习结束之后，则共享它们此时的 Q 值，作为下个周期($t+1$)学习的基础。

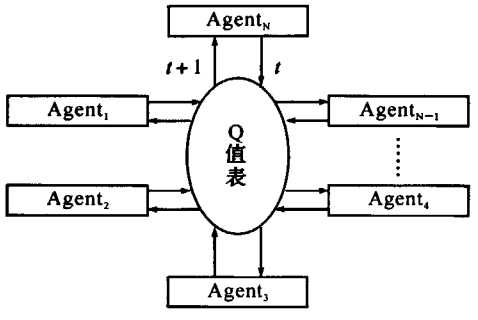

图 1　多 Agent 的知识共享模型图

3.2　学习协作算法实现

设当前 MAS 中共有 N 个 Agent，每个 Agent 每个周期都学习，用 $H(s,a)(s\in S, a\in A)$ 来表示 Agent 的某一行为的 Q 值表。$Q_{i,t}(s,a)$ 表示第 i 个 Agent 在 t 时刻的 $Q(s,a)$ 值，当它们完成一个周期的学习之后，Agent 会将其 $Q_{i,t}(s,a)$ 写入 $H(s,a)$ 中。设 $H_{i,t}(s,a)$ 表示 t 时刻第 i 个 Agent 对应的 Q 值表中的 $Q(s,a)$ 值，则多 Agent 系统学习协作算法如下：

1) 初始化各个 Agent 的 Q 函数值，将 Q 表中的初始值 Q 初始化为 0 且 $t=0$；

2) Loop until $Q(s_t,a)$ 逼近 $Q^*(s_t,a)$

End loop

$Q(s,a)$ 的迭代函数表达式为：

$$Q_{t+1}(s,a)=(1-\alpha_t)Q_t(s,a)+\alpha_t[R(s'|s,a)+\gamma Q_t(s',a')]$$

其中 α_t 为学习率，控制学习速度，γ 为折扣因子；

3) 把得到的各个 Agent 的 Q 值共享到 Q 表中，并计算在时刻 t 的所有 Agent 的 $Q(s,a)$ 的平均值 $H_{avg}^{s,a}(t)=\frac{1}{N}\sum_{i=1}^{N}H_{i,t}(s,a)$；

4) 计算各个 Agent 的奖励系数(惩罚系数) $\beta_{i,t}=\frac{H_{i,t}(s,a)}{H_{avg}^{s,a}(t)}$，$\beta_{i,t}>1$ 表示正奖励系数，$\beta_{i,t}<1$ 表示负奖励系数(惩罚系数)；

5) 更新当前状态下各 Agent 的 $Q(s,a)$，更新函数式为：$Q_{i,t}(s_t,\vec{a})=(1-\alpha_t)Q_{i,t-1}(s_{t-1},\vec{a})+\alpha_t[\beta_{i,t}R_{i,t}+\gamma\max_{a'\in A}Q_{i,t-1}(s_t,\vec{a})]$，其中 $\vec{a}$ 为联合行为向量 $\vec{a}=(a_1,a_2,\cdots,a_N)$；

6) 根据获得的知识,选择最优策略执行,更新Q值表并更新知识库;

7) $t \leftarrow t+1$,得到新状态s_{t+1}及强化信号R_{t+1};

8) 转第2)步继续执行下一状态。

上述算法的基本思想是某Agent的Q值大于其同时刻的所有Agent的平均水平时,说明该Agent此刻的Q值所反映出的状态行为能带来较好的奖励,该行为应该被加强;反之,Q值小于平均水平时,则该状态下采用的行为应该得到一定的抑制,即得到一定的惩罚,通过对其奖励或惩罚来更新Q值表。这样各个Agent在独立学习的基础上,通过对其他Agent的状态行为的比较,来调整自己的Q值,最终达到相互学习与协同动作的目的。可以证明,经过一段长时间训练学习后,整个算法能收敛到一个理想的值[6]。结合以上算法的多Agent系统的学习协作框图如图2所示。

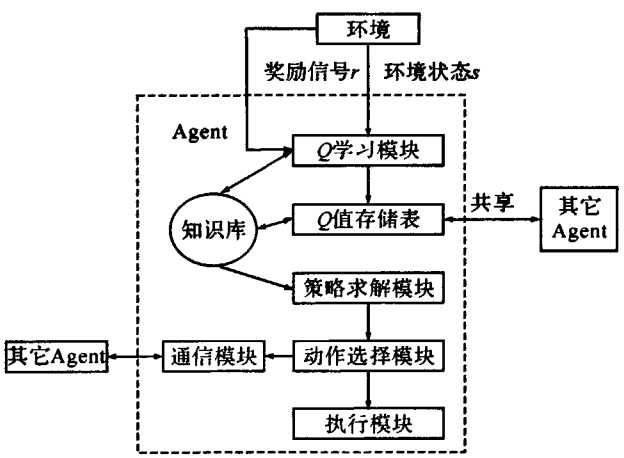

图2　多Agent学习协作框图

4　一种新的多Agent系统结构模型的提出

虽然以上提出的方法能很好的解决多Agent系统学习及协作问题,但我们也看到其中存在的一些问题:1)由于每个Agent都要参与学习,都会生成一个Q值,这样对于一个复杂的大系统来说,其存储的状态空间也要变大,随着Agent数目及环境状态的增加,Q值表的大小将成指数级递增[7]。而且其查询、搜索和计算的时间消耗也将十分可观。2)由于各Agent每个周期都要进行知识共享,这对大系统来说,这将大大影响系统的性能和效率。对此,我们考虑能不能把复杂的大系统简化成简单的小系统模型呢?

文献[8]提出了一种共享经验元组的多Agent协同强化学习算法,使得状态行为空间得到大大缩减。而在实际中,我们发现并不是每个Agent都需要学习,而只要起主导、支配作用的Agent具有学习能力即可。本文把需要学习的Agent定义为强Agent,把不需要学习的、其辅助作用的Agent定义为弱Agent,这样我们把具有相同信念的一些弱Agent和一个强Agent组成一个团队。基于以上思想,本文提出了基于团队的三层体系机构,如图3所示。

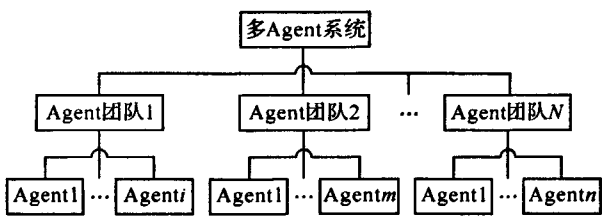

图3　多Agent系统三层体系结构

在团队中,通过一个强Agent学习,然后将经验直接分享给其余弱Agent,来保持整个团队的协同动作,而团队与团队之间应用上面的知识共享机制来保持整个多Agent系统的相互学习协作。为了减少Agent之间的相互通信,节约系统资源,结合文献[9]中提出的由一个服务商的服务Agent组成一个邦盟的思想。这里把一个团队中的强Agent看做服务商,这样其他团队的Agent发送消息到本团队某个Agent时,不是直接发送到特定的Agent,而是先发送到本团队的服务商,然后服务商将消息转发到特定的Agent。这样避免了各Agent之间的直接相连,整个Agent系统建网及通信的复杂度降低。结合以上算法,可以发现多Agent系统中Agent的维数N大大降低,Q值表的状态空间$H(s,a)$大大缩小,同时由于采用了服务商的概念,避免了传统多Agent系统通信时产生的竞争和阻塞的困扰,提高了整个系统的性能和效率,也降低了系统的建设成本。其结构模型如图4所示。

5　仿真实验

传统的基于Q-learning算法的多Agent系统把每个Agent都直接相连通信,靠它们之间复杂的通信机制来保持相互协作,并且每个任务由一个Agent来单独完成,每个Agent都需要独立学习[10]。为了验证本文提出的结构体系及机制算法比传统结构的优越性,本文将通过下面的仿真实验来进一步说明。仿真软件采用的是MATLAB 9.0。

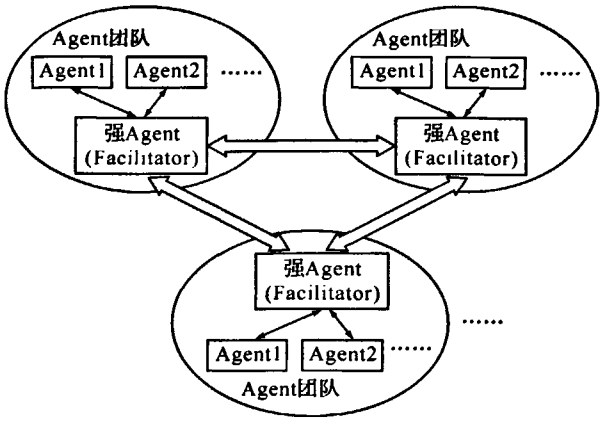

图 4　多 Agent 结构模型图

本文通过两个方面来进行验证：学习周期数和所耗费的时间步数。学习周期数表示系统达到收敛所要花费的学习周期；时间步数表示系统完成一个任务所消耗的时间数。这里我们假设 MAS 的 Agent 个数为 $N=60$，折扣系数 $\gamma=0.9$，学习率 $\alpha=0.1$，实际中奖励值随着阶段状态的变化而不断变化，为了能够模拟仿真，我们这里统一规定正的奖赏值为＋10，负的奖赏值（惩罚值）为－10；则采用上述算法后，仿真图形如图 5 所示。

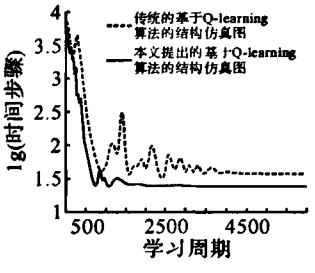

图 5　仿真波形图

从图中我们可以看出，应用新模型结构的收敛速度要比传统结构快得多，大约 1700 个学习周期后即收敛，而传统的大约在 3800 个学习周期后才收敛，这种收敛差异是减少系统维数、减少学习 Agent 个数和降低协作开销造成的。并且新模型结构收敛后达到的时间步数要少于传统模型，这是因为以前由一个 Agent 完成的任务现在由一个 Agent 团队来分担完成，试想同样的任务由 1 个 Agent 独立完成自然要比 2 个 Agent 分担的速度慢（即需要更多的时间步数才能完成）；另外新结构减少了通信的复杂度和冲突性，使通信协作变的简单有序，所费时间自然得到降低。

最后我们发现，在学习初期（前 800 个学习周期内），两者间的差异十分有限，这是因为在初期学习协作中产生的状态空间对系统的影响有限，随着学习的深入，新结构减小状态空间措施的优势得到了明显的显现。可以预知，随着多 Agent 系统中 Agent 个数 N 的增加，两者仿真曲线的差别将更加明显，新结构的优势将进一步得到证明，这里限于篇幅，本文不一一列举仿真。

6　结语

本文以强化学习 Q-learning 算法为基础，提出了多 Agent 学习协作算法，并通过进一步改进创新，具体给出了多 Agent 系统的三层体系结构和系统结构模型图，并用 MATLAB 软件进行了相应的仿真实验，仿真结果表明了该结构的优越性和巨大进步。

多 Agent 系统是近年来的一个新兴研究领域和研究热点，有着广阔的应用前景，值得我们花更多的精力去研究它。本文提出了以 Q-learning 算法为基础的一种新型多 Agent 系统的体系结构模型，但多 Agent 系统的研究还有很长的路要走，需要我们继续去研究、改进和创新，为多 Agent 系统的发展做出我们的贡献。

参 考 文 献

[1] MICHAEL Wooldridge. An Introduction To MultiAgent System[M]. New York: J. Wiley,2002

[2] 沈晶. 分层强化学习理论与方法[M]. 哈尔滨：哈尔滨工程大学出版社，2007

[3] WATKINS C, DAYAN P. Q-Learning[J]. Machine Learning,1992,8(3):279～292

[4] Bellman R E. A Markov Decision Process[J]. Journal of Mathematical Mechanics,1957,6(5):679～684

[5] 吴元斌. 单 Agent 强化学习与多 Agent 强化学习比较研究[J]. 电脑与信息技术，2009,17(1):8～11

[6] WATKINS C. Learning From Delayed Rewards[D]. Cambridge: University of Cambridge,1989

[7] 郑淑丽，韩江洪，等. 多 Agent 系统的协作及强化学习算法研究[J]. 模式识别与人工智能，2002,15(4):453～457

[8] 王长缨，尹晓虎，等. 一种共享经验元组的多 Agent 协同强化学习算法[J]. 模式识别与人工智能，2005,18(2):235～238

[9] MICHAEL R. Genesereth, STEVEN P. Ketchpel. Soft Agents[J]. Communications of the ACM,1994,37(7):48～53

[10] TAN M. Multi-agent Reinforcement Learning: independent vs. Cooperative Agents[C]//Proceedings of the Tenth International Conference on Machine Learning. [s. l.]:[s. n.],1993:330～337