改进的 CE-Q 算法用于多 Agent 觅食的研究 *

雷默涵，杨萍
（兰州理工大学 机电工程学院，甘肃 兰州 730050）

摘要：针对对策论框架下的诸多强化学习方法在复杂环境多 Agent 任务中存在的缺乏理性、难以保证收敛、计算复杂度较高和效率偏低等问题，文中在 CE-Q 强化算法的基本理论上，提出了加入对于动作过程的即时奖赏的 CE-Q 改进强化算法，有效地改善了上述问题，并在执行任务过程中对 Agent 进行指导，很好地提高了系统效率。最后以多 Agent 觅食为任务，Matlab 为平台进行仿真实试验，并与普通 CE-Q 及 FF-Q 算法进行对比，验证了其在复杂环境下对于多 Agent 系统的有效性和优越性。

关键词：CE-Q 强化学习算法；动作过程奖赏；多 Agent 觅食任务；系统效率
中图分类号：TH122 **文献标识码**：A **文章编号**：1001-2354(2015)06-0001-04
DOI:10.13841/j.cnki.jxsj.2015.06.001

Study on modified CE-Q algorithm for multi-Agent foraging

LEI Mo-han，YANG Ping
(School of Mechanical Engineering，Lanzhou University of Technology，Lanzhou 730050)

Abstract：Several reinforcement-learning algorithms under game theory were put up，but the algorithms had many problems such as lacking of rationality，non-guaranteed convergence，high calculating complexity and low efficiency in complicated environment. So，CE-Q algorithm was adopted with a reward function for actions in processes. An improved CE-Q reinforcement-learning algorithm avoided above defects and had high efficiency. Then Matlab was used to carry out simulation of foraging which was compared with normal CE-Q and FF-Q reinforcement-learning algorithms，proving that the algorithm was valid and superior for multi-agent system under complicated environment.

Key words：CE-Q algorithm；reward function for action；multi-Agent forage；efficiency of system

多 Agent（智能体）系统已经在适用性、经济性、鲁棒性、灵活性和容错性等方面表现出极大的优越性，比较适合在实际生产、生活甚至军事领域代替人力完成一些恶劣、危险环境下的工作。如果多 Agent 系统要真正发挥作用，使其自适应环境的学习控制方法是必要的。强化学习作为目前被广泛关注的机器学习方法，为多 Agent 系统自主学习提供了可行的方案。

现阶段对多 Agent 强化学习算法的研究主要集中在算法的收敛和 Agent 之间交互从而使各 Agent 兼顾自身和整体利益的问题。对策论已经为多 Agent 间的交互提供了一个可靠的数学框架，并在这个框架下提出了几种强化算法[1]。在这些算法当中，Minimax-Q[2-3]算法只能应用于两个 Agent 的零和对策之中；Nash-Q[2,4]算法可以应用于一般和对策，但算法需要使用两个非常严格的条件假设来保证 Agent 选取相同的 Nash 均衡解，且 Nash 平衡计算的复杂度较高；FFQ[2,5]算法不能兼顾 Agent 的整体理性，也不满足 Agent 的自主性；Pareto-Q 算法可以使合作的 Agent 更为理性，但不能用于非合作策略的研究，且存在最优解的计算问题。而 Amy Greenwald，Keith Hall 提出的 CE-Q[2,6]算法能够兼顾个体理性和整体理性，可以较好地保证收敛性及适用性，也可以利用线性规划求出相关均衡解，相对比较适合于实际应用中的多 Agent 任务。

多机器人觅食作为一个典型任务，在采矿、灾难救援、爆炸物处理等任务中具有广泛的应用背景[7-8]。该任务提供的是一个非结构化的复杂试验环境，这个试验环境下需要 Agent 做出一系列动作才能完成其中的一个子任务。由于传统的 Agent 奖赏方法是在 Agent 完成任务之后对其进行奖赏，只能简单地给出任务，不能有效的在这一系列动作过程当中指导 Agent 执行任

* 收稿日期：2013-09-30 修订日期：2014-12-03

务，这种情况下传统的奖赏方法会使得强化算法在试验中效率不高。基于此，文中在 CE-Q 强化算法中加入了在执行任务的过程中对 Agent 进行指导的动作过程奖赏函数，使其在任务中取得更高的效率。

1 基于即时奖赏的 CE-Q 算法

1.1 多 Agent 系统下的 Q 算法

在将单 Agent 的 Q-learning 强化算法推广到多 Agent 的协同增强学习系统时，直观的方法就是将单 Agent 的动作替换为所有 Agent 的联合动作[2,9]，因此非确定性环境下的最优 Q 值函数将被定义为：

$$Q_i^*(s,a)=r_i(s,a)+\gamma\sum_{s'}P[s,a,s']V_i^*(s') \tag{1}$$

式中：P——状态转移概率函数；

γ——折扣因子；

a——所有 Agent 联合动作集合。

在 n 个 Agent 的情况下，联合策略为 $\pi=(\pi_1,\pi_2,\cdots,\pi_n)$，最优策略定义为：

$$\pi_i^*(s)\in\arg\max_{a\in A(s)}Q_i^*(s,a) \tag{2}$$

在这种 Q 强化算法通用公式中，值函数(Value Function)V_i^*的计算方法并不是明确的。多 Agent 下的 Q 算法描述如下[2,15]：

Multi-Q

输入　折扣因子 γ

　　学习率 α

　　总时间 T

输出　最优 Q 值函数 Q_i^*

初始化　$s,a_1,\cdots,a_n$ and $Q_1,\cdots,Q_n$

For t=1 to T

　模拟动作　$a_1,\cdots,a_n$

　观察即时回报值 $r_1,\cdots,r_n$ 和下个状态 s'

　for i=1 to n

　　计算 $V_i(s')$

　　更新 $Q_i(s,a_1,\cdots,a_n)$

　　$i.\ Q_i(s,a_1,\cdots,a_n)=(1-\alpha)Q_i(s,a_1,\cdots,a_n)+\alpha[r_i+\gamma V_i(s')]$

　选择动作 $a'_1,\cdots,a'_n$

　$s=s',a_1=a'_1,\cdots,a_n=a'_n$

1.2 算法的值函数定义

针对 Q 算法中值函数 V_i^*不同的定义方法，可以将随机博弈框架下的强化学习算法分为 Minimax-Q，Nash-Q，FF-Q，CE-Q 等。CE-Q 强化学习算法是根据相关均衡解概念(Correlated Equilibrium)命名的。在这些算法中，CE-Q 算法继承了 Nash-Q 算法的优点，在策略的选择中考虑了其他 Agent 动作对某个 Agent 的影响。相关均衡解是联合动作空间上的概率分布，在这种情况下各个 Agent 将会根据其他 Agent 的动作自行调整其条件概率以最大化所有 Agent 的总回报值，实现对整个多机器人系统的优化。而且与 Nash 均衡解目前还没有有效的计算方法不同，相关均衡解可以通过对简单的线性规划问题求解得到[6-7]。

Amy Greenwald 和 Keith Hall 在对 CE-Q 强化算法值函数 V_i^*的定义中最大化了所有 Agent 的回报值之和，并称其为 utilitarian。在学习中利用线性规划的方法得到使 Agent 回报值之和最大的相关均衡解。

CE-Q 强化学习算法中值函数的具体定义如下：

$$V_i^*\in\mathrm{CE}_i(Q_1^*(s),\cdots,Q_n^*(s)) \tag{3}$$

$$\mathrm{CE}_i(Q_1^*(s),\cdots,Q_n^*(s))=\{\sum_a\sigma^*(a)Q_i(s,a)\,|\,\sigma^*\mathrm{satisfies}Eq.3\} \tag{4}$$

$$\sigma^*\in\arg\max_{\sigma\in\mathrm{CE}}\sum_{a\in A}\sigma(a)(\sum_{i\in I}Q_i(s,a)) \tag{5}$$

1.3 算法的即时奖赏函数定义

Q 强化学习算法通常是在一个子任务完成或是一种情况(例如 Agent 之间的碰撞)发生之后得到奖赏，这是一种基于结果的奖赏。但是对于较复杂环境下的任务来说，一个任务往往是由一系列的动作组成，这种奖赏函数将无法对每个动作的奖赏进行分配，从而不能指导 Agent 怎样去执行任务[8]。

基于此种情况，文中将在原有即时奖赏的基础上加入对于 Agent 动作过程的即时奖赏，以指导 Agent 执行任务：

$$r_i(t)=r_i^{\mathrm{result}}+r_i^{\mathrm{trend}} \tag{6}$$

式中：r_i^{result}——对 Agent 产生结果(Agent 捡拾到目标物等)的奖赏；

r_i^{trend}——对 Agent 动作(Agent 携带目标物时，是否向相应目标区域移动)的奖赏。

奖赏(结果与动作均发生在 t-1 时刻)情况如下：

$$r_i^{\mathrm{result}}(t)=\begin{cases}13 & \text{漫游时捡起任意目标物}\\ 20 & \text{运送目标物到目标区域}\\ -18 & \text{与Agent发生冲突}\\ -15 & \text{碰到障碍物}\\ -15 & \text{丢弃目标物}\\ 0 & \text{其他}\end{cases} \tag{7}$$

$$r_i^{\mathrm{trend}}(t)=\begin{cases}3 & \text{携带目标物时靠近目标区域}\\ -3 & \text{携带目标物时远离目标区域}\\ 0 & \text{其他}\end{cases} \tag{8}$$

1.4　改进 CE-Q 算法的实现

文中改进 CE-Q 算法是在 Q 算法的基础上，采用上述基于相关均衡解的 V_i^* 值函数及基于过程奖赏的即时奖赏函数 r_i，从而保证了算法的收敛、整体、个体理性的兼顾，并能在实际应用中实现较低的计算复杂度和较高的系统效率。

2　仿真试验验证

2.1　仿真试验平台构建

多机器人觅食作为一个多 Agent 系统研究领域的典型任务，在采矿、灾难救援、爆炸物处理等任务中具有广泛的应用背景[9]。该任务提供的是一个非结构化的复杂试验环境，具有多个子任务且每个子任务的完成都需要 Agent 完成一系列的动作，可以用于验证文中提出的强化算法是否有效。

以静态多 Agent 觅食为任务，MatLab7.0 为平台，采用基于动作奖赏的 CE-Q 算法，CE-Q 强化算法及 FF-Q 三种强化学习算法进行仿真试验。在试验中，每个 Agent 都是智能的，具有感知、移动和与每个 Agent 通信协调的能力。每个 Agent 可以在一定范围（以 r 为半径的圆形区域）里感知到环境信息，感知到的环境信息包括其他 Agent 的位置、目标物的位置、障碍物的位置和目标区域的位置，并且由于 Agent 需要联合学习合作策略，仿真试验需要假设在整个试验环境中任意 Agent 具有与环境中其他所有 Agent 通信的能力[10-11]。

仿真试验中利用手工编程设计了每个机器人的行为集合，强化学习算法通过学习得到最优联合策略 π^*，在不同状态下各 Agent 将根据最佳联合策略 π^* 选择联合动作集，获得所有 Agent 的动作控制向量，最后输出到 Agent 的运动执行单元并使 Agent 产生动作。每一次循环运动时，所有 Agent 都同时各自按照动作控制向量执行一个动作[12]。

图 1a 为 6 个 Agent 觅食任务的仿真环境，试验环境中还包括 6 个矩形障碍物和 5 个圆形障碍物，灰、黑两个目标区域，10 个灰色目标物和 10 个黑色目标物。由于 CE-Q 算法存在维度灾难问题，试验中对多种环境状态进行了聚类，为减少动作向量空间，在试验中假设机器人只向上、下、左、右 4 个方向运动，且 Agent 数量不多于 8 个。

2.2　仿真试验过程

在试验中 Agent 将会随机漫游寻找目标物，判断目标物的种类然后将其收集并运送进相应的目标区域。仿真试验将采用静态环境，即目标物、障碍物和目标区域都静止不动。仿真试验中首先采用了基于动作过程奖赏的 CE-Q 算法，然后采用 CE-Q 算法及 FF-Q 算法，最后通过分析比较验证基于即时奖赏的 CE-Q 算法的有效性和优越性。

仿真试验中，当 Agent 之间相撞或者碰到障碍物时，Agent 将会退回原处，且即时事件奖赏 $r_{i\,\mathrm{result}}^{a}(t)$ 为 -1。所有障碍物皆为不可通过，Agent 必须选择避开障碍物才可通过（当 Agent 不携带相应目标物时，也需要避开目标区域）。Agent 开始处于漫游状态，当 Agent 在感知范围内发现目标物的存在时，其捡拾状态就会被触发，Agent 将前往捡起目标物。随后 Agent 的状态转移成为运送态，成功将目标送到目标区域后，状态将会再次变为漫游态，继续出发寻觅目标物。

最后所有 Agent 收集到目的地的灰、黑目标物总量能很直观地反应系统性能，数量越多代表系统性能越好。每种算法将在同样的环境下进行循环试验，采用折扣因子 γ=0.8，学习因子 α_n 初始为 0.5，将随着迭代次数的增加逐渐变小，变化形式为[13]：

$$\alpha_n=\frac{0.5}{1+\mathrm{visits}_n(s,a)}\quad(0\leqslant\alpha_n<1)$$

式中：$\mathrm{visits}_n(s,a)$——n 次循环内被访问的次数。

2.3　仿真试验结果及分析

图 1b 为加入过程动作即时奖赏的 CE-Q 算法应用于 6 个 Agent 的静态觅食任务试验。可以看到所有 Agent 都避开了障碍物，没有发生相互之间的冲突。此时灰色目标区域中已存放了 32 个目标物，黑色目标区域中则有 30 个目标物。

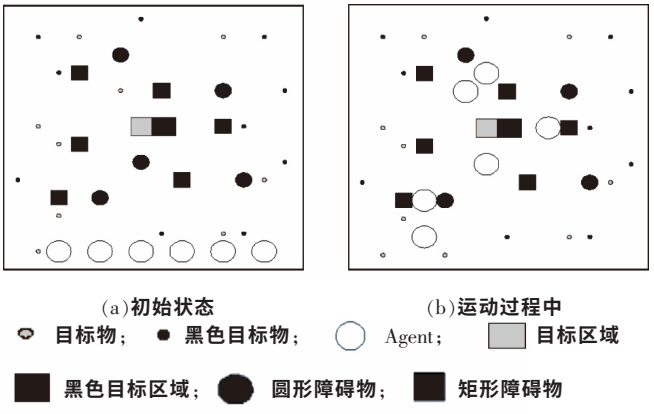


图 1　6 个机器人觅食活动仿真试验环境

图 2 是 3 种算法在觅食任务中的收敛情况，基于动作过程奖赏的 CE-Q，CE-Q 及 Friend-Q 三种算法均可以收敛。图 3 为系统性能曲线图，可以看出，加入

过程动作即时奖赏的 CE-Q 算法收敛速度快于其他两种强化算法，并且在不同 Agent 数量下收集到的灰色、黑色目标物均多余其他两种算法。

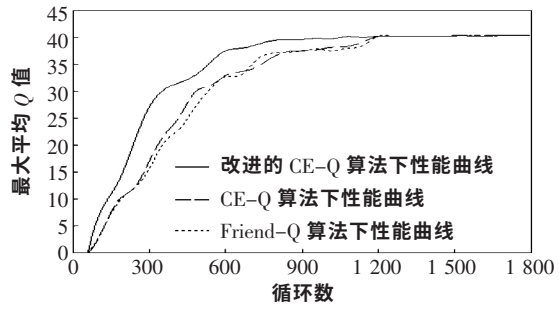
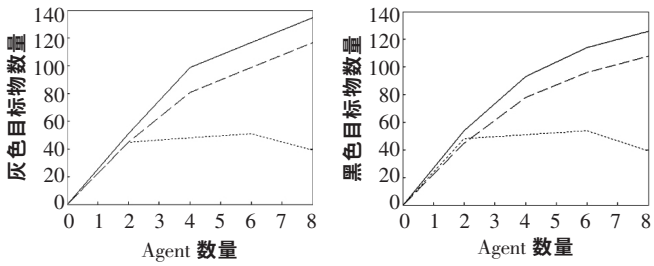


图 2 3 种算法在觅食任务中全部收敛



(a)不同 Agent 数量时收集到的灰色目标物数量

(b)不同 Agent 数量时收集到的黑色目标物数量

—— 改进的 CE-Q 算法下性能曲线； —— CE-Q 算法下性能曲线；
····· Friend-Q 算法下性能曲线

图 3 系统性能曲线图

当环境中 Agent 数量增加时，收集到的目标物数量也会增加，但是由于 Agent 之间的冲突也随之加剧，系统性能反而会下降。Friend-Q 由于 Agent 之间无法协调各自的策略选择，在 Agent 数量较多时表现不佳，Agent 数目超过 2 个后对系统性能的提升微弱，到达 8 个时，收集到目标区域的目标物数量开始下降，1800 次循环后仅仅收集到 39 个灰色目标物和 39 个黑色目标物。CE-Q 算法表现正常，但由于不能在动作过程中指导 Agent 执行任务，系统性能在不同的 Agent 数量时系统的效率均逊色于加入过程动作即时奖赏的 CE-Q 算法，在 8 个 Agent 系统中只收集到 108 个灰色目标物和 117 个黑色目标物，而加入过程动作即时奖赏的 CE-Q 算法收集到 126 个灰色目标物和 138 个黑色目标物。

3 结语

针对多 Agent 强化学习算法在实际任务中表现出缺乏理性、适用性差、计算复杂度高等问题，提出加了动作过程奖赏的 CE-Q 改进强化算法，且以 Matlab 仿真为试验手段，在多 Agent 觅食任务中与普通 CE-Q 及 FF-Q 算法进行对比，试验表明加入动作过程奖赏的 CE-Q 强化算法有以下优点：

(1)继承了 CE-Q 算法均衡解计算复杂度低、适用范围较广的优势。

(2)通过对 Agent 在执行任务过程中的动作行为进行指导，提高了多 Agent 系统的效率，使其具有更快的收敛速度和更好的系统性能。

参考文献

[1] 吴军，徐昕，王健，等. 面向多机器人系统的增强学习研究进展综述[J]. 控制与决策，2011，26(11)：134-138.

[2] 宋国平，顾国昌，张国印. 随机博弈框架下得多 Agent 强化学习方法综述[J]. 控制与决策，2005，20(10)：67-73.

[3] Michael L Littman. Markov games as a framework for multi-Agent reinforcement learning[C]//11th Int Conf on Machine Learning，New Brunswick，1994：157-163.

[4] Hu J L，Michael P Wellman. Multi-Agent reinforcement learning：theoretical framework and an algorithm[C]//15th Int Conf on Machine Learning，Madison，1998：242-250.

[5] Michael L Littman. Friend-or-foe Q-learning in general-sum markov games[C]//18th Int Conf on Machine Learning，MA：Williams College，2001：322-328.

[6] Greenwald A，Keith Hall. Correlated Q-learning[C]//Proc of the 20th Int Conf on Machine Learning，Madison，1998：242-250.

[7] 宋伟科. 基于多机器人的开放式智能控制系统关键技术研究与开发[D]. 天津：天津大学，2012.

[8] 任燚，陈宗海. 基于强化学习算法的机器人系统[D]. 安徽：中国科学技术大学，2005.

[9] 姜新丽. 基于强化学习的多机器人协作控制方法研究[D]. 沈阳：沈阳理工大学，2011.

[10] 余江，杨威，费燕群. 混合型自重构机器人的空间构型描述[J]. 机械设计，2013，30(8)：12-14.

[11] 刘海江，姜冬冬，张春伟. 面向室内场景的空地多机器人协作环境感知[J]. 机械设计，2011，28(6)：33-16.

[12] 刘璇，张明路，张小俊，等. 基于公理设计的模块化特种机器人配置过程研究[J]. 机械设计，2011，28(2)：40-45.

[13] 王雪松，程玉虎. 机器学习理论、方法及应用[M]. 北京：科学出版社，2009.

[14] 徐昕. 增强学习与近似动态规划[M]. 北京：科学出版社，2010.

[15] Tom M. Mitchell. 机器学习 [M]. 北京：机械工业出版社，2003.

作者简介：雷默涵(1987—)，男，硕士研究生，研究方向：特种环境下的工业机器人，已发表论文 2 篇。E-mail：13919174135@163.com
杨萍(1964—)，女，教授，硕士生导师，博士，研究方向：特种环境下的工业机器人、CAD/CAM 及虚拟设计，发表论文 60 余篇。