基于改进蚁群算法的 Q 学习算法研究

付 鹏,罗 杰
(南京邮电大学 自动化学院,江苏 南京 210046)

摘 要:文中以围捕问题作为研究平台,以提高多 Agent 系统中 Q 学习算法的学习效率作为研究目标,提出了一种基于改进蚁群算法的 Q 学习算法。该算法将信息素的概念引入到 Q 学习中,结合采用动态自适应调整信息素挥发因子的蚁群算法,使 Agent 在进行行为决策时不再只以 Q 值作为参考标准,而是考量 Q 值与信息素的综合效应,加强了 Agent 彼此间的信息共享,增强了交互性。并且对于复杂变化的周围环境,根据具体环境条件,设立分阶段的多奖惩标准,使算法对于环境和状态有更好的适应性。仿真实验证明了改进后的 Q 学习算法提高了学习系统的效率,高效地实现了多 Agent 系统的目标任务。

关键词:多 Agent 系统 Q 学习;改进蚁群算法;围捕问题
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2013)02-0123-04
doi:10.3969/j.jssn.1673-2013.02.031

Q Learning Algorithm Research Based on Improved Ant Colony Algorithm

FU Peng,LUO Jie
(College of Automation,Nanjing University of Posts & Telecommunications,Nanjing 210046,China)

Abstract:On the basis of round up problem,propose a Q-learning algorithm based on the improved ant colony algorithm in order to improve the learning efficiency of Q-learning algorithm in the multi-Agent system. The algorithm introduces the concept of pheromone to Q-learning process,combined with ant colony algorithm of dynamic adaptive adjustment volatile factor,so that the Q value is no longer the only determinant for decision-making of the Agent,but is considered to be a part of the combined effect with pheromone,which can strengthen the sharing of information and enhance the interaction between Agents. In addition,the Q-learning algorithm can set phased incentive standards according to complex surrounding environment and specific conditions,achieve better adaptability of itself to the environment and conditions. The simulation experience results show that the improved Q-learning algorithm increases the efficiency of the learning system and achieves objectives of the multi-Agent system effectively.

Key words:Q-learning algorithm in the multi-agent system;improved ant colony algorithm;round up problem

0 引 言

强化学习作为一种重要的机器学习,是近年来众多研究者关注的前沿领域和热点课题。它是一种以环境反馈作为输入的机械学习方法,采用这种机制的 Agent 通过与周围环境的交互来实现对最优行为策略的学习[1]。

但是在多 Agent 系统中,传统的 Q 学习算法具有其自身的局限性。当系统中 Agent 的个数增加,或某种状态下 Agent 可选择的动作增加时,会导致整个状态空间呈指数上升,进而使得动作选择的策略和搜索速度均不理想,引起学习效率极速下滑,带来“维数灾难”问题。在文中,以围捕问题作为平台来研究多 Agent 系统的 Q 学习算法,在方格世界中,四名猎人在不断地学习与协作下来捕获一个猎物,通过模拟最具有智能性的人类思维方式来实现 Agent 的学习与合作[2]。以提高系统中 Agent 学习效率为目标改进学习算法,文中通过将信息素的概念引入到多 Agent 系统中,结合了一种采用动态自适应调整信息素挥发因子的改进蚁群算法,增强了各 Agent 间的交互。此外,针对不同环境,根据具体需要,制定了不同的奖惩方式,避免采用单一、固定的联合奖惩办法。通过上述改进,在一定程度上解决了维数灾难,提高了学习效率。在具体的围捕研究平台上体现为猎人捕获猎物的成功率

收稿日期:2012-06-03;**修回日期**:2012-09-06
基金项目:江苏省高校自然科学基金项目(04KJB110097,08KJB520023)
作者简介:付 鹏(1987-),男,硕士,研究领域为智能机器人、模式识别与人工智能等;罗 杰,博士,教授,研究领域为分布式智能控制、群体智能。

提高,捕获猎物所需的时间下降。

1 改进的 Q 学习算法

1.1 Q 学习算法

在与模型无关的学习算法中,Q 学习算法是非常重要的算法,它是基于马尔科夫决策过程的递增式动态规划算法。马尔科夫决策过程由五元组组成,对于此种决策过程,Agent 进行的每一步,均需要先确认当前所处的状态 s_t,然后参照某种控制策略,从动作集合中选择动作 a_t 执行,之后转移到下一个状态,并得到一个即时回报。在 Q 学习算法中,每个 $Q(s_t,a_t)$ 都有与之相应的一个 Q 值,它就是按照所选择的策略持续执行而得到的累积回报[3]。

Q 学习算法中累计回报定义如下[4]:

$$Q(s_t,a_t)\rightarrow(1-\alpha)Q(s_t,a_t)+\alpha[r+\gamma\max Q(s_{t+1},a_{t+1})] \tag{1}$$

其中 α 为学习率($\alpha>0$),γ 为折扣因子($0\leqslant\gamma<1$)。在该算法中,任一状态 s_t 下最优策略所采取的动作,就是选择具有最大回报值的动作。

1.2 基于改进蚁群算法的 Q 学习算法

文中将蚁群算法中信息素的概念引入,通过每个 Agent 在环境中留下各自的标记信息来实现各个 Agent 间信息的合作与共享,增强了交互,具体思想为:每次当 Agent 在环境中进行学习时,都会在所处环境中留下自己的信息素,记为 τ,当某处 τ 的量越多时,越说明经过该处 Agent 越多,就越可能是最优解路径,这样就实现了 Agent 间信息的彼此交互。

随着时间的推移,新的信息素不断增加,而旧的信息素则逐渐淡化,甚至消失。按照基本蚁群算法公式,对信息素进行如下更新[5]:

$$\begin{aligned}&\tau_{ij}(t+1)=(1-\rho)\cdot\tau_{ij}(t)+\Delta\tau_{ij}(t)\\&\Delta\tau_{ij}(t)=\sum_{k=1}^{n}\Delta\tau_{ij}^{k}(t)\end{aligned} \tag{2}$$

其中,$\tau_{ij}(t)$ 表示 t 时刻在坐标(i,j)处信息素量大小,$\Delta\tau_{ij}(t)$ 表示 t 时刻(i,j)处信息素增量,$\tau_{ij}^{k}(t)$ 表示 t 时刻 Agent k 在(i,j)处信息素增量;ρ 表示信息素挥发系数,$1-\rho$ 表示信息素残留因子。

对于 ρ,传统的蚁群算法将其定义为[0,1)范围内的一个常数,而 ρ 的大小直接关系到蚁群算法的全局搜索能力和收敛速度[6],所以在这里采用改进的算法,让信息素挥发因子 ρ 按下式作自适应调整:

$$\rho(t)=\begin{cases}0.95\rho(t-1) & 0.95\rho(t-1)\geqslant\rho_{\min}\\ \rho_{\min} & \text{否则}\end{cases} \tag{3}$$

式中,$\rho_{\min}$ 为 ρ 的最小值,是为了防止 ρ 过小而降低算法的收敛速度。

传统的 Q 学习算法选择的动作策略是选取最大回报值的动作执行,也就是说只通过参考 Q 值来决定下一步的行为动作。有了上面所引入的信息量,采用下面策略公式:

$$\pi_{s_t}=\operatorname{argmax}p_{ss'}[\tau_{ij}+\gamma Q(s',a)] \tag{4}$$

其中,$p_{ss'}$ 为执行动作从状态 s 到状态 s' 的转移概率,γ 为折扣因子。这样就使得一个 Agent 考虑到其它 Agent 的信息,进行决策时不再单依赖 Q 值作为依据,而是考量 Q 值与相应信息量的综合效应来进行动作执行。如果某一个动作的 Q 值比较大,可是信息素少,或者说两者综合起来的评价要差于其它的动作,那么 Agent 就选择总体评价高的那步动作,而不再仅仅去选择 Q 值高的那个。

文中还根据猎人 Agent 是否到达目标区域,采用了分情况的多奖惩办法,也就是说这里将有多个状态集。其好处在于 Agent 能够在不同环境条件下,先进行判断行为,然后在选定的条件中再进行下一步动作,这样就有效减少了 Agent 的搜索空间,对搜索速度的提高带来促进,进而缓解因为联合动作和联合状态而带来的维数灾难问题[7]。

算法具体如下:

第一步:初始化每个 Agent 的状态集,动作集,初始化 $Q(s,a)$ 表,设置各项参数;

第二步:重复执行操作:

1 观察 t 时刻的状态 s;

2 按照公式(4) 策略执行动作 a;

3 执行所选动作,并观察下一个状态,判断所处的任务阶段和所处环境,选择适合该状态的奖惩标准,得到回报值 r,同时在相应位置处留下信息素;

4 根据公式更新 Q 值和信息素;

5 判断是否满足终止条件,满足则停止学习,不满足继续执行第二步。

2 基于改进蚁群算法的 Q 学习算法的围捕问题

2.1 围捕问题模型

作为一个经典的人工智能问题,猎人围捕问题模型被很多研究人员用来研究多 Agent 系统的学习、协作问题。文中,背景环境是一个 7×7 的方格世界,其中有四个 Agent 猎人,一个 Agent 猎物,整体实验界面环境如图 1 所示(图中表示的为猎人 Agent 围捕到猎物 Agent 时的情形)。

图中星形表示猎人,圆表示猎物,猎物被放置在中心,猎人则在其它位置随机分布,将每一个方格代表一种状态,猎人 Agent 有如下五种动作:上、下、左、右和静止不动。

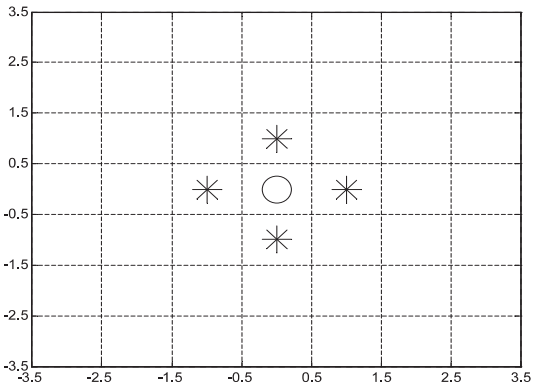

图 1　围捕模型示意图

对围捕模型增加一个信息储存机制,用来存储信息素,即每个 Agent 走过某个位置时,环境能够感应并记录相应信息。当 Agent 处于某个位置,就根据坐标记录下其在此位置处的信息以及接下来移动的方向,当 Agent 向某个方向移动越多,该方向的信息素量就越大,例如,一个 Agent 走到某个位置的时候,会在该处留下一定的信息素,如果下一步向左走,那么相应方向的信息素就会增加一些。这样每个 Agent 通过感应各处信息素的量来指导学习,从而增强了各个 Agent 间的交互,使其更好地学习。换句话说,当某处某方向的信息素越浓,Agent 向某处移动的概率会越高,就会越趋向最优解路径,最终实现捕获任务。

在学习过程中,猎人不能同时到达同一位置,若试图到达同一位置则被退回原处[8]。在开始围捕任务前,四名 Agent 猎人会根据其自身的坐标信息来选择目标位置:纵坐标最大猎人的负责猎物的北边位置;纵坐标最小猎人的负责猎物南边位置;横坐标最大的猎人负责猎物的东边位置;横坐标最小的负责猎物的西边位置。围捕成功标志为:猎人 Agent 都到达目标位置。

文中采用分阶段式的奖惩办法,根据具体情况采用不同的奖惩标准。以猎物为原点建立坐标系,然后用一、三象限和二、四象限的 45 度线将这个坐标系划分为四个部分。当猎人 Agent 和目标位置在同一区域时,利用横纵坐标控制猎人 Agent 与猎物 Agent 的距离变化来实现奖惩规则;反之,则根据其实际位置与期望位置间的坐标关系建立奖惩标准。所以在这里有两个状态集,一个负责到达目标区域前的奖惩规则,另一个负责到达目标区域后的奖惩规则。

总的来说,猎人围捕的策略是先形成包围圈,然后再逐渐缩小包围圈直至捕获。期间通过综合参考各处留下的信息素量和相应的奖惩规则来选择较好的移动路径,最终有效地实现目标任务。

2.2　仿真实验及结果分析

文中算法实验参数为:$\alpha = 0.1$, $\gamma = 0.95$, $\varepsilon = 0.8$;根据相距目标"变远、不变、变近"设置 $r = \{-1, 0, 1\}$;初始 Q 值设为 0,各处的信息素量初始也为 0;信息素挥发因子的初始值 $\rho(0) = 1$。以一百步次作为移动上限,任何一个猎人 Agent 超过此上限步次,均设定为任务失败。猎人 Agent 的初始位置会随机分布,每一次的学习都在上一次实验基础上继续进行[9],实验任务一共进行一百次,图 2 为第一阶段实验结果:

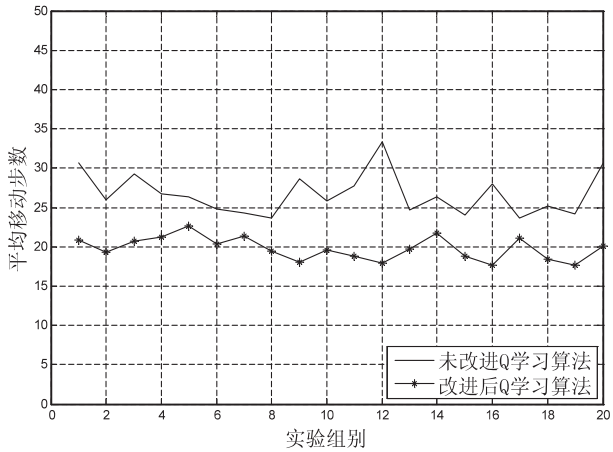

图 2　两种算法性能比较

实验采用对比方式,第一阶段里,分别对改进前后的算法各进行一百次实验,一共进行二十个统计小组,记录全部猎人 Agent 在成功实验中的步数。根据图线变化明显看出,基于改进蚁群算法的 Q 学习总体趋势波动更小,走向更为平稳,所需要的平均步数大致减少七步左右。总之,图 2 的实验结果体现了基于改进蚁群算法的 Q 学习算法的稳定性[10]。

第二阶段就两种算法的性能进行具体对比分析,在实验中,第一轮采用没有改进的标准 Q 学习算法,第二轮采用改进后的 Q 学习算法。设定每个小组负责五次实验,则一百次实验共分二十个统计小组完成,得到的二十组实验数据为每个统计小组围捕任务成功的次数。实验结果如图 3、图 4 所示:

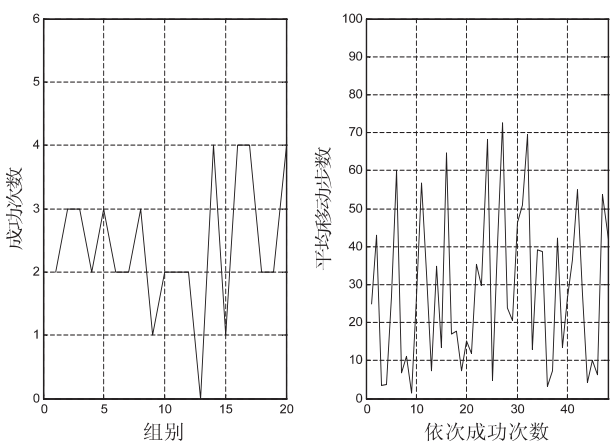

图 3　标准 Q 学习算法实验结果

图 3、图 4 左边图线的意义表示各个统计小组记录的相应围捕任务成功次数,而在右边图中,通过统计

每次实验围捕任务成功时,猎人 Agent 行动的步次,最后计算出所有猎人 Agent 行动的步次平均值,并显示在图线上,即右边的图线中,横轴记录了捕获任务成功的次序,竖轴显示全部猎人 Agent 在相应的成功任务下行动步次的平均值[11]。

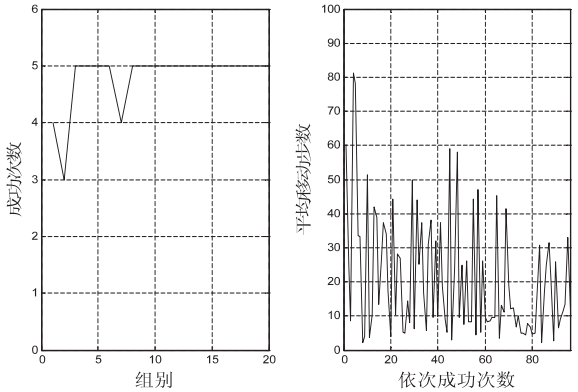

图 4 改进后 Q 学习算法试验结果

根据结果,从曲线整体走向上比较分析标准 Q 学习算法与基于改进蚁群算法的 Q 学习算法[12]。首先分析左图,可以看出算法改进前后,在整体上都令围捕任务成功次数得到一定增加,而新的算法这种上升趋势更为明显。而从两组图的右边图线可以看到,未改进前的标准 Q 学习算法成功了大约 50 多次,而改进后的新学习算法则成功了 80 多次,成功次数明显增加。另外,依然从走势上看,原来的标准算法中的图线没有很明显的趋势变化,而改进后的新算法则随着成功次数的增加,总体上有一个较为明显的下降趋势,即随着成功次数的增加,完成捕获任务所需的步数有所下降。

3 结束语

文中以围捕问题为研究平台,提出多 Agent 系统中基于改进蚁群算法的 Q 学习算法,采用动态自适应调整信息素挥发因子,并将信息素的概念引入系统中,通过考量 Agent 的 Q 值和所处位置信息素的综合效应来决定行为策略。此外面对多变的环境,还采用了符合实际情况的多奖惩标准,这使得文中的改进算法在弥补缺陷,整体提高算法收敛性上都取得一定效果。

仿真实验结果很好地证明了基于改进蚁群算法的 Q 学习算法能够更好地适应所处环境,有效提高学习效率[13]。

参考文献:

[1] Watkins C J C H, Dayan P. Technical note: Q-learing[J]. Machine Learning, 1992, 8(3-4): 279-292.

[2] 孟祥萍,王圣镔,王欣欣. 多 Agent Q 学习几点问题的研究及改进[J]. 计算机工程与设计,2009,30(9):2274-2276.

[3] 刘 杰. 基于强化学习的多机器人围捕策略的研究[D]. 长春:东北师范大学,2009.

[4] 黄炳强. 强化学习方法及其应用研究[D]. 上海:上海交通大学,2007.

[5] 赵凤遥,廖宏骞. 基于改进蚁群算法的优化设计[J]. 科学技术与工程,2009,9(19):5902-5905.

[6] 贾瑞玉,张新建,冯伦阔,等. 信息素增量动态更新的改进蚁群算法[J]. 计算机技术与发展,2009,19(9):32-34.

[7] 周浦成,洪炳熔,黄庆成. 一种新颖的多 Agent 强化学习方法[J]. 电子学报,2006,34(8):1488-1491.

[8] 张 林,徐 勇,刘福成. 多 Agent 系统的技术研究[J]. 计算机技术与发展,2008,18(8):80-83.

[9] 胡子婴. 基于智能体系统的 Q-学习算法的研究与改进[D]. 哈尔滨:哈尔滨理工大学,2007.

[10] 贺建民,闵 锐. 多 Agent 系统中蚁群算法的设计与实现[J]. 微电子学与计算机,2006,23(10):32-34.

[11] 段海滨. 蚁群算法原理及其应用[M]. 北京:科学出版社,2006.

[12] 马 勇,李龙澍,李学俊. 基于 Q 学习的 Agent 智能防守策略研究与应用[J]. 计算机技术与发展,2008,18(12):106-108.

[13] Ito K, Imoto Y, Taguchi H. A study of reinforcement learning with knowledge sharing[C]//Proceedings of the 2004 IEEE International Conference on Robotics and Biomimetics. Japan: Okayama University Digital Information Repository, 2004: 175-180.

(上接第 122 页)

[12] Fatourechi M, Bashashati A, Ward R K, et al. EMG and EOG artifacts in brain computer interface systems: a survey[J]. Clinical Neurophysiology, 2007, 118(3): 480-494.

[13] Schlogl A, Kronegg J, Huggins J E, et al. Evaluation criteria in BCI research[M]//Toward brain-computer interfacing. [s. l.]: MIT Press, 2007: 327-342.

[14] Schlogl A, Keinrath C, Zimmermann D, et al. A fully automated correction method of EOG artifacts in EEG recordings[J]. Clin. Neurophys, 2007, 118(1): 98-104.

基于改进蚁群算法的Q学习算法研究

作者： 付鹏， 罗杰
作者单位： 南京邮电大学 自动化学院,江苏 南京 210046
刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2013(2)