在线更新的信息强度引导启发式 Q 学习*

吴昊霖[1,2], 蔡乐才[2,3], 高 祥[2]
(1. 四川理工学院 自动化与信息工程学院, 四川 自贡 643000; 2. 人工智能四川省重点实验室, 四川 自贡 643000; 3. 宜宾学院, 四川 宜宾 644000)

摘 要: 针对强化学习收敛速度慢的问题,提出可在线更新的信息强度引导的启发式 Q 学习算法以加快强化学习算法的收敛速度。该算法在启发式强化学习算法的基础上引入依据每次训练回报情况进行在线更新的信息强度,通过结合强弱程度不同的动作信息强度更新的启发函数和状态—动作值函数来确定策略,从而提高算法收敛速度。给出该算法并对其收敛性进行证明,同时针对不同参数设置和仿真环境进行路径规划的仿真对比实验,得到实验结果:信息强度引导的启发式 Q 学习算法在成功率、达到目标位置所需步数及所获回报上均优于 Q 学习和基本启发式 Q 学习算法。实验结果表明,该算法能更快地得到回报较高的策略且不会陷入局部收敛,因而该算法能够有效提高算法的收敛速度。

关键词: 强化学习; 启发函数; 信息强度; 在线更新; 收敛速度
中图分类号: TP181 **文献标志码**: A **文章编号**: 1001-3695(2018)08-2323-05
doi:10.3969/j.issn.1001-3695.2018.08.021

Online pheromone stringency guiding heuristically accelerated Q-learning

Wu Haolin[1,2], Cai Lecai[2,3], Gao Xiang[2]
(1. *School of Automation & Information Engineering, Sichuan University of Science & Engineering, Zigong Sichuan* 643000, *China*; 2. *Artificial Intelligence Key Laboratory of Sichuan Province, Zigong Sichuan* 643000, *China*; 3. *Yibin University, Yibin Sichuan* 644000, *China*)

Abstract: Since reinforcement learning is time-consuming algorithm, this paper presented an online pheromone stringency guiding heuristically accelerated Q-learning algorithm to speed up the convergence rate. Based on the heuristically accelerated reinforcement learning, heuristic function added a pheromone stringency which could be updated online according to the rewards of the training, then combined with the value function to determine the policy. This paper proved the convergence of the algorithm. Meanwhile, simulation results show that pheromone stringency guiding heuristically accelerated Q-learning algorithm has better performance in the rate of success, steps to reach target location and rewards obtained in the learning. The algorithm can find the optimal policy faster and avoid getting into local convergence, thus effectively speeding up the convergence rate.

Key words: reinforcement learning(RL); heuristic function; pheromone stringency; online; convergence rate

强化学习(RL)是机器学习重要分支之一。不同于监督学习(supervised learning)和无监督学习(unsupervised learning),它通过试错(trial-and-error)的方式与环境进行交互以完成学习。若环境对其动作评价为积极的,则选择该动作的趋势加强,否则便会减弱。Agent 在不断训练的过程中根据环境的评价得到最优策略[1]。因此强化学习具有自主学习和在线学习的特点,在近几年受到越来越多的关注,也得到越发广泛和复杂的实际应用[2~4]。

虽然强化学习有着诸多优点以及值得期待的应用前景,但强化学习也存在着收敛速度慢、维数灾难、平衡探索与利用、时间信度分配等问题[5]。强化学习收敛速度慢的原因之一是没有教师信号,只能通过探索并依靠环境评价逐渐改进以获得最优动作策略[6]。为进一步加快强化学习的收敛速度,启发式强化学习通过给强化学习注入一定的先验知识,有效提高强化学习的收敛速度。Torrey 等人[7]通过迁移学习为强化学习算法注入先验经验以提高收敛速度;但是迁移学习所注入的先验知识是固定的,即使有不合理规则也无法在训练过程中在线修正。Bianchi 等人[8]通过给传统强化学习算法添加启发函数,在训练过程中结合使用值函数和启发函数来选择动作,提出了启发式强化学习(heuristically accelerated reinforcement learning, HARL)算法模型。启发式强化学习最重要的特点是在线更新启发函数,以不断增强表现更好的动作启发函数。方敏等人[9]在启发式强化学习算法的基础上提出一种基于状态回溯的启发式强化学习方法,通过引入代价函数描述重复动作的重要性,结合动作奖赏及动作代价提出一种新的启发函数定义以进一步提高收敛速度,但是该方法只是针对重复性动作的重要性进行评估。

本文在启发式强化学习算法的基础上借鉴蜂群采蜜过程中通过摇摆舞持续时间表示蜜源质量的想法,提出信息强度引导的启发式 Q 学习(pheromone stringency guiding heuristically accelerated reinforcement learning, PSG-HAQL)算法,进一步提高启发式强化学习的收敛速度。

1 启发式 Q 学习

Q 学习是较为常用的强化学习算法,学习方式依然为在环境中不断试错,在没有环境模型的条件下通过自举(bootstrapping)的方式更新状态—动作值函数,寻找最优策略。Q 学习中状态—动作值函数的更新规则如式(1)所示。

$$Q(s_t,a_t)=Q(s_t,a_t)+\alpha[R+\gamma\max_a Q(s_{t+1},a)-Q(s_t,a_t)] \quad (1)$$

其中:α 为学习率,通常来说学习率越大则收敛速度越快,但较易产生振荡;γ 为折扣因子,表示较近动作的重要程度;动作

收稿日期: 2017-03-31; **修回日期**: 2017-05-15 **基金项目**: 国家自然科学基金资助项目(61202196);人工智能四川省重点实验室项目(2015RYY02);企业信息化与物联网测控技术四川省高校重点实验室开放基金项目(2014WZY01,2016WYJ02);四川理工学院研究生创新基金资助项目(Y2016032)

作者简介:吴昊霖(1990-),男,山东临沂人,硕士,主要研究方向为强化学习、迁移学习(wuhaolin@126.com);蔡乐才(1966-),男,四川大竹人,教授,博士,主要研究方向为机器学习、计算机网络安全;高祥(1983-),男,山东青岛人,讲师,硕士,主要研究方向为机器学习、智能仪器仪表.

$a \in A$,A 为动作空间;状态 $s_t, s_{t+1} \in S$,S 为状态空间;R 为 agent 由当前状态执行动作到达下一状态所获得的奖励值;$Q(s_t, a_t)$ 为动作值函数。

式(1)中 Q 学习采用最大化的策略去更新动作值函数,属于离策略(off-policy)的强化学习方法。离策略强化学习的特点是用于评估和改进策略的评估策略(estimation policy)与用于产生动作的行为策略(behavior policy)不一致。在 Q 学习算法中,评估策略是确定性策略;而行为策略可采用随机策略去遍历所有的状态。其中,选择动作的随机行为策略对 Q 学习算法的效率有很大的影响[10]。

启发式 Q 学习通过启发函数,与 Q 学习算法结合起来作用于行为策略,从而影响 Q 学习算法的动作选择以加快学习速度。启发式 Q 学习算法(HAQL)的动作选择机制如下所示:

$$\pi(s) = \begin{cases} \arg\max_a [Q(s_t, a) + H(s_t, a)] & \text{if } q < \beta \\ a_{\text{random}} & \text{otherwise} \end{cases} \tag{2}$$

在启发式 Q 学习中,启发函数 $H(s_t, a)$ 的更新公式如下:

$$H(s_t, a) = \begin{cases} \max_a Q(s_t, a) - Q(s_t, a_t) + \eta & \text{if } a_t = \pi^H(s_t) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

其中:$\pi^H(s_t)$ 是指启发函数 H 下的最优动作[8]。

2 信息强度引导的启发式 Q 学习

仿生学为科学技术的发展提供诸多新思路,蚁群算法、蜂群算法等群集智能算法已得到广泛应用。研究表明,蚂蚁和蜜蜂是通过传递信息完成种群整体智能行为最成功的两种群居昆虫。得益于信息素,蚁群算法和蜂群算法已经成为近年来最有效的优化算法之一[11]。不同于蚁群根据随时间挥发的信息素累积情况表示信息强弱,蜂群信息特点是蜜蜂通过摇摆舞传递食物源信息,且摇摆舞的持续时间与食物源的收益率呈正比。

2.1 信息强度的更新方法

强化学习中,agent 由初始状态转移到终止状态的过程称为情节(episode),或译为幕。其中终止状态包括认定为成功的目标状态和认定为失败的避免状态。在 agent 到达目标终止状态时会得到一个收益,结合蜂群信息随食物源的质量呈正比变化的思想,将每一情节的收益作为食物源的收益率称为适应度 f。若判断该情节学习所获得的适应度 f 大于之前的最大适应度 $f_{\max}$,则根据所获得适应度大小更新该情节学习所遍历状态的信息强度。设计表 H 来记录与适应度 f 呈正比关系的信息强度,同时定义信息强度的更新规则用于启发强化学习的探索与利用。

定义 1 设计表 H 记录与适应度呈正比的信息强度,其元素为四元组$\langle s_i, a_i, p(s_i, a_i), f_{\max}\rangle$。其中:$s_i$ 为需要更新信息强度的信息状态;a_i 为需要更新信息强度的信息动作;$p(s_i, a_i)$ 为更新后的信息强度;$f_{\max}$ 为此前记录的信息状态 s_i 适应度最大值。信息强度 $p(s_i, a_i)$ 的更新规则如式(4)所示。

$$p(s_i, a_i) = \begin{cases} \dfrac{p(s_i, a_i) f_{\max}}{f} & \text{if } a_i \neq a_t \\ \Delta & \text{if } a_i = a_t \end{cases} \tag{4}$$

其中:a_t 表示 agent 最新情节的学习中在状态 s_i 采用的动作;a_i 表示表 H 中的信息动作;Δ 默认值为 1,随后随着信息强度的衰减而更新;$f_{\max}$ 表示表 H 中的适应度最大值;f 表示训练过程中新获得的适应度,其更新方式为

$$f = \sum_{s=s_0}^{s_t} (\beta R) \tag{5}$$

通过以上更新规则,使信息强度 $p(s_i, a_i)$ 由适应度 f 与表 H 中适应度最大值 $f_{\max}$ 的差值程度所决定。当 f 大于表 H 中储存的 $f_{\max}$ 时,信息强度则需要更新,即表 H 需要更新。基于上述更新规则,该算法在保留此前信息强度的同时,使按照适应度差值程度更新的信息强度体现出不同信息动作的重要性。

2.2 信息强度的启发方法

当前强化学习动作选择策略主要有 Boltzmann 和 ε-贪婪策略两种机制。目前还无法确定两种策略选择方法哪个更好,只是根据任务不同而进行选择[12]。当采用式(3)所示的 ε-贪婪策略选择动作时,由于叠加启发函数后总会使启发函数下最优动作的混合函数最大,从而只选择信息素强度下最优动作;而在非贪婪选择动作情况下对每个动作的选择是随机的,即表现最差的动作也有同等的概率被选择。Boltzmann 机制根据动作表现好坏而选择概率不同。

信息强度引导的启发式 Q 学习的动作选择策略采用 Boltzmann 机制,其更新方式规则如式(6)所示。

$$P(a_i | s) = \frac{e^{[Q(s,a_i) + H(s,a_i)]/T}}{\sum_{k=1}^{N} e^{[Q(s,a_k) + H(s,a_k)]/T}} \tag{6}$$

当采用 Boltzmann 机制,若当前最大动作值函数下的动作不是信息素强度下的最优动作,则通过 $Q(s_t, a) + H(s_t, a)$ 加大信息素强度下最优动作的选择概率;同时 Boltzmann 机制在不同动作信息素强度差距不大的情况下,使得最大动作值函数下的动作与信息素强度下最优动作的概率相近,从而避免陷入信息素强度下的局部最优;在信息素强度差距较大的情况下,使动作选择概率偏向于信息素强度下最优动作,从而有助于算法收敛。此外,Boltzmann 机制使其他动作也有一定概率被选择,从而促进算法探索。

在 PSG-HAQL 算法中,上述信息强度将直接影响启发式强化学习的启发函数 $H(s_t, a)$,进而影响强化学习中的动作选择策略。在 HAQL 算法基础上对式(2)所示的动作选择策略进行改进;同时为了突出信息强度的影响,进一步加快启发强化学习收敛速度,本文中定义一种新的启发函数。

定义 2 为使所获得的信息强度大小直接反映在动作选择上,将信息强度融入到启发函数。通过设置影响量级参数来控制信息强度对动作选择的影响程度。启发函数更新方式定义为

$$H_t(s_t, a) = \begin{cases} \max_a Q(s_t, a) - Q(s_t, a_t) + \dfrac{p(s_t, a_t)}{\text{sum}(p(s_t, a))} U & \text{if } a_t = \pi^p(s_t) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

其中:$\pi^p(s_t)$ 是在信息强度启发下的最优动作;$\dfrac{p(s_t, a_t)}{\text{sum}(p(s_t, a))} U$ 是通过最大信息强度与信息强度总和比重来表示该动作的重要性,记为 h;U 是信息强度对动作选择的影响量级参数,U 越大则信息强度的影响越大。

在以上更新规则中,只有信息强度启发下最优动作的启发函数进行更新,作用于动作策略的选择,非信息素强度启发下最优动作的启发函数均被设为 0。当信息素强度启发下最优动作的值函数小于另一动作时,通过叠加启发函数使动作选择更加倾向于信息素强度较大的动作,而不是在不完全探索情况下选择值函数较大的动作。注意,如式(6)所示,启发函数并不是直接作用于动作值函数,使动作值函数发生变化;而是进行叠加操作,将叠加函数用于决定动作选择策略,继而此情节学习的回报作用于动作值函数的更新。

2.3 信息强度的衰减机制

在信息强度引导的启发式强化学习中,由于 agent 的策略选择是在启发函数和值函数的叠加作用下确定的,并在策略动作之后更新值函数。当信息强度引导的启发式强化学习算法趋于稳定时,引入信息强度的衰减机制,通过信息强度的衰减逐渐减小启发函数的影响量级。

定义 3 当信息强度引导的启发函数连续 M 情节的训练不发生更新时,进行信息强度的自衰减以减少启发函数的影响量级。信息强度衰减机制的更新规则如式(8)所示。

$$p(s_i, a_i) = \frac{\max_a Q(s_i)}{\max_a Q(s_i) + H(s_i, a_i)} p(s_i, a_i) \tag{8}$$

通过以上更新规则,使得信息强度的衰减程度与启发函数和状态—动作值函数的相对比值有关。当启发函数比状态—

动作值函数大很多时,信息强度衰减程度比较大;反之衰减程度比较小。最终使得启发函数的影响逐渐减小,主要通过动作—状态值函数确定策略。

2.4　PSG-HAQL 算法

初始化表 $H\langle s_i, a_i, p(s_i,a_i), f_{\max}\rangle$
初始化动作值函数 $Q(s_t,a)$
初始化启发函数 $H(s_t,a)$
初始化 $\Delta=1$
repeat
　初始化状态 $s\leftarrow s_0$
　repeat
　　根据式(6)在状态 s_t 选择动作 a_t

$$P(a_i|s)=\frac{e^{[Q(s,a_i)+H(s,a_i)]/T}}{\sum_{k=1}^{N} e^{[Q(s,a_k)+H(s,a_k)]/T}}$$

　　执行动作 a_t,获得回报 R 和下一状态 s_{t+1}
　　根据式(1)更新动作值函数

$$Q(s_t,a_t)=Q(s_t,a_t)+\alpha[R+\gamma\max_a Q(s_{t+1},a)-Q(s_t,a_t)]$$

　　根据式(5)更新适应度值 $f=\sum_{s=s_0}^{s_t}(\beta R)$
　　更新状态 $s\leftarrow s_{t+1}$
　until 到达终止状态
　if $f\geqslant f_{\max}$
　　信息强度衰减标志置零
　　根据式(4)更新 H 表:

$$p(s_i,a_i)=\begin{cases}\dfrac{p(s_i,a_i)f_{\max}}{f} & \text{if } a_i\neq a_t\\ \Delta & \text{if } a_i=a_t\end{cases}$$

　　更新最大适应度 $f_{\max}=f$
　　根据式(7)更新启发函数:

$$H(s_t,a)=\begin{cases}\max\limits_a Q(s_t,a)-Q(s_t,a_t)+\dfrac{p(s_t,a_t)}{\mathrm{sum}(p(s_t,a))}U & \text{if } a_t=\pi^p(s_t)\\ 0 & \text{otherwise}\end{cases}$$

　else
　　信息强度衰减标志执行自加操作
　if 满足信息强度衰减条件

$$p(s_i,a_i)=\frac{\max\limits_a Q(s_i)}{\max\limits_a Q(s_i)+H(s_i,a_i)}p(s_i,a_i)$$

$$\Delta=\frac{\max\limits_a Q(s_i)}{\max\limits_a Q(s_i)+H(s_i,a_i)}\Delta$$

　until 停止条件满足

3　PSG-HAQL 算法收敛性分析

PSG-HAQL 算法中采用信息强度作为启发函数,即当 $a_t=\pi^H(s_t)$ 时启发函数为

$$H(s_t,a)=\max_a Q(s_t,a)-Q(s_t,a_t)+\frac{p(s_t,a_t)}{\mathrm{sum}(p(s_t,a))}U$$

该模型仍属于广义的马尔可夫模型,其最优策略不变性和 Q 值迭代收敛性已被证明[13~15]。这里还需针对其在信息强度启发下动作选择策略的收敛性进行证明。

证明　假设在状态 s^*,其初始最优策略动作为 a,在训练过程中执行动作 a_2 获得更大的适应度值 f^*,则此时根据式(4)更新信息素强度,即

$$p(s^*,a_1)<p(s^*,a_2)$$

则

$$\pi^p(s^*)=\max_a p(s^*,a)=a_2$$

根据式(7),在 $a=a_2$ 的情况下:

$$H(s^*,a_2)=\max_a Q(s^*,a)-Q(s^*,a_2)+\frac{p(s^*,a_2)U}{\mathrm{sum}(p(s^*,a))}$$

在 $a=a'$的情况下,其中 a'为包括 a_1 在内不等于 a_2 的动作:

$$H_t(s^*,a')=0$$

根据式(6),在动作选择策略中:

$$Q(s^*,a_2)+H(s^*,a_2)=$$

$$Q(s^*,a_2)+\max_a Q(s^*,a)-Q(s^*,a_2)+\frac{p(s^*,a_2)U}{\mathrm{sum}(p(s^*,a))}=$$

$$\max_a Q(s^*,a)+\frac{p(s^*,a_2)U}{\mathrm{sum}(p(s^*,a))}$$

其中:$\dfrac{p(s^*,a_2)U}{\mathrm{sum}(p(s^*,a))}>0$,又

$$Q(s^*,a')+H(s^*,a')=Q(s^*,a')+0=Q(s^*,a')$$

显然

$$\max_a Q(s^*,a)+\frac{p(s^*,a_2)U}{\mathrm{sum}(p(s^*,a))}>Q(s^*,a')$$

即　　$Q(s^*,a_2)+H(s^*,a_2)>Q(s^*,a')+H(s^*,a')$

则在信息素强度影响量级参数 U 较大时,即在利用阶段,根据式(6)有

$$\pi(s^*)=a_2$$

由此,证明在信息素强度启发下动作选择策略收敛在信息素强度大的策略;且通过已被证明的最优策略不变性和 Q 值迭代收敛性,PSG-HAQL 算法收敛于最优策略。

4　仿真实验和结果分析

为验证 PSG-HAQL 算法的有效性,将该算法应用于机器人路径规划问题,并通过在相同环境下分别采用 PSG-HAQL 算法、高强度启发式 Q 学习算法(H-HAQL)、低强度启发式 Q 学习算法(L-HAQL)和基本 Q 学习算法(Standard-QL)作为对比进行仿真实验。

4.1　实验环境及参数设置

为排除实验结果偶然性,分别在 20×20 仿真环境和 30×30 仿真环境进行对比实验。20×20 仿真环境如图 1 所示,实心红色区域(见电子版)代表障碍物,四周代表墙壁障碍;起点位置设置为(1,1),目标位置设置为(18,17)。30×30 仿真环境如图 2 所示,起点位置设置为(1,1),目标位置设置为(29,20)。

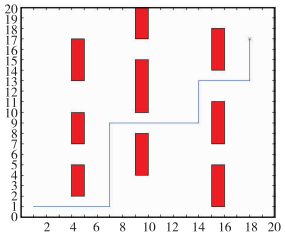


图1　20×20仿真实验

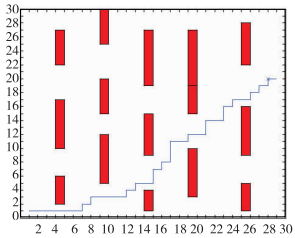


图2　30×30仿真实验

目标位置回报值为 100,其余位置回报值均按照该位置与目标位置的距离差的大小分布在(0,2]内,距离差越小则回报值越大。为防止 agent 为获得回报而产生不必要的移动,将回报值设置为负值,即回报值的范围设置为[-2,0)的标量,该位置与目标位置的距离差越小,回报值依然越大。Agent 动作空间为{1,2,3,4},分别代表向上、向下、向左、向右。若 agent 撞到障碍物或者墙壁,则退回起点,并得到-10 的惩罚。

在采用不同方法进行仿真实验时,均设置为相同参数,如表 1 所示。为尽可能保证实验结果准确,对每种方法分别进行 20 次实验,其中 20×20 仿真环境每次实验的 Episode 设置为 3 000,30×30 仿真环境每次实验的 Episode 设置为 4 000。取该 20 次实验的数据均值的四舍五入近似值作为实验结果进行分析。其中,PSG-HAQL 的信息强度影响量级参数设置为 1.5;HAQL 为文献[8]中的启发式 Q 学习,H-HAQL、L-HAQL 的 η 分别设置为 1.5、0.1,用于与 PSG-HAQL 作对比实验;Standard-QL 为标准的 Q 学习算法。

表 1　仿真实验参数设置

实验次数 n	学习率 α	折扣因子 γ	恒定温度 T_1	动态温度 T_2
20	0.5	0.9	0.5	0.8→0.2

使用强化学习算法时通常采用动态参数去平衡算法的探索与利用,即在开始阶段促使算法倾向于探索,而经过一定的训练后通过参数的动态调节使算法更倾向于利用。在四种算法的对比实验中,为避免动态参数影响实验结果,首先采用恒

定温度参数观察实验结果。为防止强化学习算法过于探索或者过于利用,将强化学习算法的温度参数 T 参数设置为0.5。然后再采用动态温度参数进行实验观察实验结果。动态温度参数设置如下:随着训练过程中成功率的变化,动态温度参数由0.8向0.2递减。

4.2 实验结果及分析

为避免实验结果的偶然性,在上述仿真环境以及参数设置下共进行四组实验,即20×20仿真环境恒定温度参数实验、20×20仿真环境动态温度参数实验、30×30仿真环境恒定温度参数实验和30×30仿真环境动态温度参数实验。在每一组仿真实验中,分别采用PSG-HAQL、H-HAQL、L-HAQL、Standard-QL算法进行对比实验。

此外,本文给出以下三个参数描述实验结果:

a)学习过程累计成功率。到达目标位置的学习情节数与学习情节总数的比值。

b)每情节学习所用步数。该情节学习找到目标位置所用的步数;如果没有到达目标地点,则步数为0。

c)每情节学习所获得累计回报值。该情节学习从起始状态到达终止状态(障碍物或者目标位置)所获得的累计回报值。

其中,学习过程累计成功率提供针对四种对比算法性能的总体认识,但其只是针对每情节学习是否到达目标位置进行统计,并不能直接由此判定每情节学习四种算法效果。每情节学习所用步数可以直观展示在每情节的学习过程中agent策略的优劣,但在环境较为简单的情况下,四种算法均能找到最优策略,所以增设每种算法在学习中使用最优策略到达目标点的次数。此外,agent可通过不同路径到达目标位置,不同路径所需步数大多不同,但也有可能不同路径的步数相同;为此设置参数每情节学习所获得累计回报值。

四组仿真实验结果的主要数据如表2~5所示。

表2 20×20环境恒定温度参数实验结果

算法	Episode	Suc/%	Iter	Step	Num	Rew
PSG-HAQL	3 000	81.1	1 021	33	752	-12.74
H-HAQL	3 000	78.5	1 175	37	/	-13.19
L-HAQL	3 000	76.7	1 299	35	/	-12.97
Standard-QL	3 000	75.5	1 383	39	/	-14.44

表3 20×20环境动态温度参数实验结果

算法	Episode	Suc/%	Iter	Step	Num	Rew
PSG-HAQL	3 000	80.3	1 163	33	1 469	-12.74
H-HAQL	3 000	77.9	1 313	33	2	-12.85
L-HAQL	3 000	75.1	1 471	33	20	-12.75
Standard-QL	3 000	74.6	1 499	33	1	-12.77

表4 30×30环境恒定温度参数实验结果

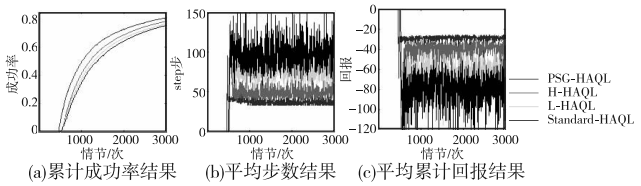
算法	Episode	Suc/%	Iter	Step	Num	Rew
PSG-HAQL	4 000	77.6	1 693	47	277	-35.67
H-HAQL	4 000	75.1	1 823	55	/	-44.95
L-HAQL	4 000	67.2	2 403	55	/	-42.64
Standard-QL	4 000	65.8	2 549	73	/	-57.18

表5 30×30环境动态温度参数实验结果

算法	Episode	Suc/%	Iter	Step	Num	Rew
PSG-HAQL	4 000	67.3	2 417	47	594	-35.67
H-HAQL	4 000	61.5	2 813	53	/	-40.65
L-HAQL	4 000	56.3	3 401	51	/	-39.13
Standard-QL	4 000	55.8	3 545	73	/	-53.02

表中,Episode为每次仿真实验的最大训练情节数;Suc为每次仿真实验的总成功率;Iter为当总成功率达到50%所需要的训练情节数,用于定量描述成功率的增长情况;Step为经过训练agent所确定的最优策略下到达目标点所需要的步数;Num表示每种算法在训练中使用最低步数到达目标点的次数,用于在Step相等的情况下定量对比四种算法性能;Rew为经过训练agent所确定的最优策略下到达目标点所获得的回报值。

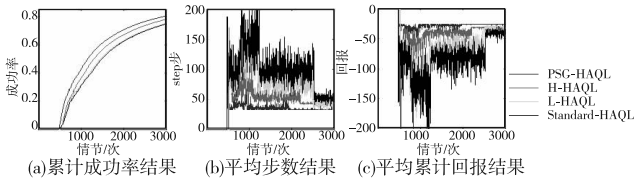
在20×20的仿真环境中,使用恒定的温度参数 T 的对比实验中,如图3(a)所示,PSG-HAQL算法的成功率增长最快,且最后的成功率也最高。如表2所示,PSG-HAQL、H-HAQL、L-HAQL算法到达50%总成功率所需的迭代次数分别比Standard-QL算法的少362、208、84次,总成功率对比趋势也大体相同,由此验证采用启发式强化学习能够帮助agent尽快地找到目标位置。此外,如表2所示,PSG-HAQL、H-HAQL算法的上述迭代次数远少于L-HAQL、Standard-QL算法,由此推断启发函数的影响量级能够提高启发式强化学习的性能。如图3(b)(c)所示,在实验所设定的Episode训练中,PSG-HAQL、H-HAQL振荡幅度比较小,而L-HAQL、Standard-QL振荡幅度较大,远未收敛;总体表现上PSG-HAQL、H-HAQL的确定策略的步数和确定策略的回报值的表现也更好一些。如表2所示,PSG-HAQL在确定最优策略的步数和回报值上均明显高于其他算法,表现为PSG-HAQL确定最优策略的步数为33步、回报值为-12.74,其余三种算法的步数和回报值均没有达到该标准,由此说明最优策略在Episode次的训练中确定的最优策略优于其他三种算法;另外,H-HAQL确定最优策略的步数和回报值分别为37和-13.19,而L-HAQL的分别为35和-12.97,H-HAQL的表现反而不如L-HAQL的,由此推断H-HAQL由于其较高的启发函数,虽然能够帮助其在训练中尽快收敛,但是也很有可能导致算法陷入局部收敛。



(a)累计成功率结果 (b)平均步数结果 (c)平均累计回报结果

图3 20×20仿真环境—恒定温度参数仿真实验结果

在强化学习算法中,常使用动态参数方法来调节算法的探索与利用问题以提高强化学习算法的性能。在20×20的动态温度参数仿真对比实验中,如表3和图4(a)所示,四种算法的总成功率有一定程度的下降,这是因为在训练前期温度参数较大,使得算法更倾向于探索,随之成功率降低;除此之外,四种算法的成功率趋势与20×20恒定温度参数仿真对比实验中的趋势大体相同,PSG-HAQL、H-HAQL、L-HAQL算法到达50%总成功率所需的迭代次数分别比Standard-QL算法的少336、186、28次。间接验证启发函数的影响量级能够提高启发式强化学习的性能,即启发函数的影响量级较大时,启发式强化学习能够更快地寻找到目标位置。当采用动态温度参数时,如表3所示,四种算法均能在设定的Episode下获得步数最少的策略,但是PSG-HAQL、H-HAQL、L-HAQL、Standard-QL在训练中使用最低步数到达目标点的次数依次为1 469次、2次、20次和1次,显然PSG-HAQL算法更能够收敛到最优策略。此外,通过表3和表2的对比,验证使用动态参数方法的确能提高强化学习的性能表现。值得注意的是,在动态参数仿真对比实验中,如表3所示,L-HAQL、Standard-QL算法所确立最优策略的回报值已优于H-HAQL算法;如图4(b)(c)所示,训练后期H-HAQL算法已达到基本收敛,但是其性能表现不如其他三种算法。这说明较大的启发函数量级虽然能使agent更快寻找到目标位置,增加成功率,但是也会导致算法陷入局部收敛。如表3和图4(b)(c)所示,PSG-HAQL算法最终也趋于收敛,且其收敛的策略明显优于其他三种算法。所以,通过信息强度的引导更新启发函数影响量级的PSG-HAQL算法,能够有效提高算法收敛到最优策略的速度。

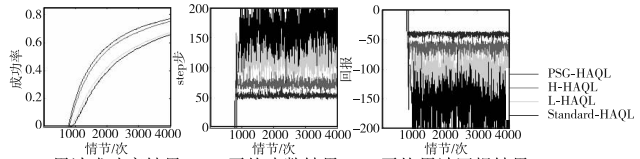


(a)累计成功率结果 (b)平均步数结果 (c)平均累计回报结果

图4 20×20仿真环境—动态温度参数仿真实验结果

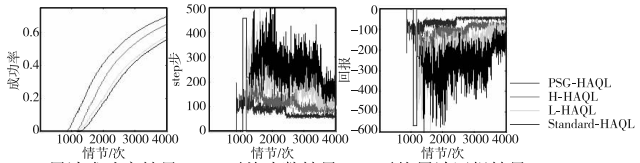
为避免实验的偶然性,将仿真环境扩展到30×30,并且增

加障碍的复杂度,如图 2 所示。由于环境变复杂,所以同时将恒定温度参数的仿真实验和动态温度参数的仿真实验的 Episode 设置为 4 000。在恒定温度参数仿真实验中,如图 5(a)所示,PSG-HAQL、H-HAQL 算法的成功率远远高于 L-HAQL、Standard-QL;同样地,如表 4 所示,H-HAQL 和 L-HAQL 虽然均获得步数为 55 的最优策略,但是 L-HAQL 确定最优策略的回报值要低于 H-HAQL。此外,只有 PSG-HAQL 算法在设定的 Episode 下寻找到全局最优策略。且随着环境的复杂化,PSG-HAQL 算法的性能改善更加明显。在动态温度参数仿真实验中,如表 5、图 6(b)(c)所示,在较为复杂的环境中,只有 PSG-HAQL 算法在设定 Episode 次训练中能够找到最优策略,其他三种算法虽然较之恒定温度参数仿真实验结果有所改进,但是在设定 Episode 次训练中均未能找到最优策略。其他实验结果与 20×20 仿真环境中的大体相同。



(a)累计成功率结果　(b)平均步数结果　(c)平均累计回报结果

图5　30×30仿真环境—恒定温度参数仿真实验结果



(a)累计成功率结果　(b)平均步数结果　(c)平均累计回报结果

图6　30×30仿真环境—动态温度参数仿真实验结果

通过以上四组仿真对比实验,首先验证了随着每次训练结果优劣情况更新的启发函数能够改善强化学习性能表面。虽然启发函数的影响量级较大时能够使 agent 更早地找到目标位置,但是较大的启发函数影响量级也会使算法陷入局部收敛。采用依据实验结果优劣更新的信息强度,然后由信息强度更新启发函数的影响量级来启发强化学习的 PSG-HAQL 算法能够更早地收敛到最优策略,有效提高强化学习算法的收敛速度。

5　结束语

启发式 Q 学习算法可以在线更新启发函数,加快强化学习速度。本文在启发式 Q 学习的基础上,提出了可在线更新的信息强度引导的启发式 Q 学习(PSG-HAQL)算法。在本文中,首先给出了信息强度更新方法和信息强度启发方法的定义,同时应用 Boltzmann 机制进行启发式强化学习,进而给出了 PSG-HAQL 算法;然后证明了 PSG-HAQL 算法的收敛性;最后将 PSG-HAQL 算法应用于路径规划仿真实验以验证其性能。PSG-HAQL 算法将蜂群信息传递的思想结合到启发式 Q 学习方法:agent 在训练过程中不断获得不同策略的适应度以在线更新该策略信息强度,将信息强度作为 Q 学习启发函数,使 agent 有更高概率去选择信息强度高的策略。所以,信息强度引导的启发式 Q 学习(PSG-HAQL)算法能够更高效地寻找到最优策略,从而进一步缩减训练时间。本文的研究工作主要集中于将蜂群信息思想引入到启发式 Q 学习模型中,并通过仿真实验验证其性能改善;下一步的工作主要是考虑将信息强度引导的启发式强化学习与价值网络相结合,从而在一定程度上提高深度 Q 网络(DQN)的收敛速度。

参考文献:

[1] 王雪松, 朱美强, 程玉虎. 强化学习原理及其应用[M]. 北京: 科学出版社, 2014: 1-2.

[2] Mnih V, Kavukcuoglu K, Silver D, *et al*. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518 (7540): 529.

[3] Levine S, Wagener N, Abbeel P. Learning contact-rich manipulation skills with guided policy search[C]//Proc of IEEE International Conference on Robotics and Automation. Piscataway, NJ: IEEE Press, 2015: 156-163.

[4] 陈学松, 杨宜民. 强化学习研究综述[J]. 计算机应用研究, 2010, 27(8): 2834-2838,2844.

[5] 于俊. 模型无关的贝叶斯强化学习方法研究[D]. 苏州: 苏州大学, 2014.

[6] 张汝波. 提高强化学习速度的方法研究[J]. 计算机工程与应用, 2001, 37(22): 38-40.

[7] Torrey L, Shavlik J, Walker T, *et al*. Skill acquisition via transfer learning and advice taking[M]. Berlin: Springer, 2006: 425-436.

[8] Bianchi R A, Ribeiro C H, Costa A H. Accelerating autonomous learning by using heuristic selection of actions[J]. Journal of Heuristics, 2008, 14(2): 135-168.

[9] 方敏, 李浩. 基于状态回溯代价分析的启发式 Q 学习[J]. 模式识别与人工智能, 2013, 26(9): 838-844.

[10] Sutton R S, Barto A G. Reinforcement learning: an introduction [M]. 2nd ed. Cambridge, MA: MIT Press, 2016: 138-139.

[11] 冀俊忠, 魏红凯, 刘椿年, 等. 基于引导素更新和扩散机制的人工蜂群算法[J]. 计算机研究与发展, 2013, 50(9): 2005-2014.

[12] 胡晓辉. 一种基于动态参数调整的强化学习动作选择机制[J]. 计算机工程与应用, 2008, 44(28): 29-31.

[13] Bianchi R A C, Riberiro C H C, Costa A H R. Heuristically accelerated *Q*-learning: a new approach to speed up reinforcement learning [C]//Advances in Artificial Intelligence. Berlin: Springer, 2004: 245-254.

[14] 魏英姿, 赵明扬. 强化学习算法中启发式回报函数的设计及其收敛性分析[J]. 计算机科学, 2005, 32(3): 190-193.

[15] 刘全, 高阳, 陈道蓄, 等. 一种基于启发式轮廓表的逻辑强化学习方法[J]. 计算机研究与发展, 2008, 45(11): 1824-1830.

(上接第 2310 页)

[2] 冀俊忠, 刘志军, 刘红欣, 等. 蛋白质相互作用网络功能模块检测的研究综述[J]. 自动化学报, 2014, 40(4): 577-593.

[3] Luo Feng, Yang Yunfeng, Chen C F, *et al*. Modular organization of protein interaction networks [J]. Bioinformatics, 2007, 23(2): 207-214.

[4] Spirin V, Mimy L A. Protein complexes and functional modules in molecular networks[J]. Proc of the National Academy of Sciences, 2003, 100(21): 12123-12128.

[5] 王杰, 梁吉业, 郑文萍. 一种面向蛋白质复合体检测的图聚类方法[J]. 计算机研究与发展, 2015, 52(8): 1784-1793.

[6] 胡赛, 熊慧军, 李学勇, 等. 多关系蛋白质网络构建及其应用研究[J]. 自动化学报, 2015, 41(12): 2155-2163.

[7] 刘阳, 季新生, 刘彩霞. 网络社区发现优化: 基于随机游走的边权预处理方法[J]. 电子与信息学报, 2013, 35(10): 2335-2340.

[8] 汪涛, 刘阳, 席耀一. 一种基于核心节点跳转的局部社区发现算法[J]. 上海交通大学学报, 2015, 49(12): 1809-1816.

[9] Wang Jianxin, Ren Jun, Li Min, *et al*. Identification of hierarchical and overlapping functional modules in PPI networks [J]. IEEE Trans on Nanobioscience, 2012, 11(4): 386-393.

[10] Palla G, Derenyi I, Farkas I, *et al*. Uncoverring the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435(7043): 814-818.

[11] Liu Guimei, Wong L, Chua H N. Complex discovery from weighted PPI networks[J]. Bioinformatics, 2009, 25(15): 1891-1897.

[12] lasblom J, Wodak S J. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs[J]. BMC Bioinformafics, 2009,10 (1): 99.

[13] Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and hierarchical community structure in complex networks[J]. New Journal of Physics, 2009, 11(3): 033015.

[14] Xenarios I, Salwinski L, Duan X J, *et al*. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions[J]. Nucleic Acids Research, 2002, 30(1): 303-305.