DOI: 10.19650/j.cnki.cjsi.J2006140

基于个体-协同触发强化学习的多机器人行为决策方法*

徐雪松[1,2],曾 智[2],邵红燕[2],杨胜杰[1],李 想[1]
(1. 湖南工商大学计算机与信息工程学院 长沙 410205; 2. 新零售虚拟现实技术湖南省重点实验室 长沙 410205)

摘 要:为了提高多机器人行为最优决策控制中强化学习的效率和收敛速度,研究了多机器人的分布式马尔科夫建模与控制策略。根据机器人有限感知能力设计了个体-协同感知触发函数,机器人个体从环境观测结果计算个体-协同触发响应概率,定义一次触发过程后开始计算联合策略,减少机器人间通讯量和计算资源。引入双学习率改进 Q 学习算法,并将该算法应用于机器人行为决策。仿真实验结果表明,当机器人群组数量在 20 左右时,本文算法的协同效率较高,单位时步比为 1.085 0。同时距离调节参数 η 对机器人协同搜索效率有影响,当 $\eta=0.008$ 时,所需的移动时步比和平均移动距离都能达到最小值。通过双学习率的引入,该算法较基于环境模型的强化学习算法具有更高的学习效率和适用性,平均性能提升 35%,对于提高多机器人自主协同能力具有较高的理论意义及应用价值。

关键词:多机器人;强化学习;个体-协同触发;行为决策
中图分类号:TH89 TP242.6 **文献标识码**:A **国家标准学科分类代码**:510.8050

Multi-robot behavior decision making method based on individual-collaborative trigger reinforcement learning

Xu Xuesong[1,2], Zeng Zhi[2], Shao Hongyan[2], Yang Shengjie[1], Li Xiang[1]
(*1.College of Computer and Information Engineering, Hunan University of Technology and Business, Changsha 410205, China;*
2.Key Laboratory of Hunan Province for New Retail Virtual Reality Technology, Changsha 410205, China)

Abstract: In order to improve the efficiency and convergence speed of reinforcement learning in multi-robot behavior optimal decision making control, the distributed Markov modeling and control strategy for multi-robots are studied in this paper. According to the limited perception ability of the robots, an individual-cooperative trigger perception function is designed. The individual robot calculates the individual-cooperative trigger response probability from the environment observation results, and defines that after a trigger process the joint strategy calculation starts, which reduces the communication amount and computing resources among robots. The Q-learning algorithm is improved through introducing the dual-learning rate strategy, which is applied to the behavior decision-making of robots. The simulation experiment results show that the algorithm proposed in this paper has quite high cooperative efficiency when the number of robots in the group is about 20. The unit time step ratio is 1.085 0. At the same time, the distance adjustment parameter η has an influence on the cooperative search efficiency of the robot. When η is 0.008, the required moving time step ratio and average moving distance can reach minimum. Through introducing the double learning rate, the proposed algorithm possesses higher learning efficiency and applicability compared with the reinforcement learning algorithm based on environment model, the average performance improvement reaches about 35%. The proposed algorithm has a high theoretical significance and application value for improving the autonomous cooperative ability of multi-robots.

Keywords: multi-robot; reinforcement learning; individual-collaborative triggering; behavioral decision-making

收稿日期:2020-03-02 Received Date:2020-03-02
*基金项目:国家自然科学基金重大项目(71991463,71790615)、国家自然科学基金重大研究计划集成项目(91846301)、湖南省教育厅科学研究重点项目(18A303)、湖南社科基金项目(18YBA272)、湖南省社科评审委员会项目(XSP18YBZ123)、湖南省重点实验室开放研究基金项目(18-07)资助

0 引　　言

随着多机器人技术在工业控制、商业服务、物流、灾害救险等领域的广泛应用,多机器人系统对复杂和不确定环境的适应性能力及协同效率成为重要研究领域[1-2]。多机器人在执行具体任务过程中,需要完成环境感知、行动规划、群体通信、综合决策等多目标多任务的协作,为此,机器人对环境的感知、识别、判断并作出相应调节的能力,是多机器人协同控制及算法研究的关键[2-9]。事实上,多机器人系统的学习是一个复杂的交互学习过程,外部环境知识获取困难且难以建模。与单个机器人相比较,多机器人系统在设计控制策略的时候,通常难以设定各个机器人所有的最优行为。虽然基于行为的方法能够让多机器人系统完成比较复杂的任务,但是仅采用基于行为的方法还不能完全适应不断变化的外界环境和不同任务的需求。让多机器人系统具有自主的学习能力,提高个体之间的协调协作能力是多机器人控制策略研究的重要发展方向[10-11]。多机器人系统间的信息协商、分配及感知能力的差异也是导致基于模型和专家知识的传统方法难以取得好的控制效果[12]。近年来,随着强化学习技术的快速发展,其良好的试错和延时回报机制,为多智能体系统自适应外界环境变化提供了可行方案[13-14]。其中 Q 学习方法[15-18]作为一典型的强化学习方法,已被广泛地应用于机器人领域。多机器人系统的个体与环境交互扩展至分布式局部可观测马尔可夫决策过程,在其学习框架中,一般分为联合动作学习体和独立学习体[18-21]。联合动作学习的计算维度会随着智能体个数增加及动作维度的膨胀,引起维数灾难问题;独立学习中各智能体动作状态独立,且无须智能体间的通信,学习模式简单,但没有充分利用各智能体环境的学习经验,复杂环境下系统的学习能力难以提升[21-22]。文献[23]结合多智能体并行采样和学习经验复用机制,提出最优值保留的复用算法提供学习经验和效率;文献[24]设计了复杂网络环境的局部通信分布式 Q 算法,提高了多机器人系统强化学习收敛性;文献[25]提出一种基于规则的关系型强化学习方法,减少机器人所用的勘探时间。多机器人个体与环境的交互及学习过程具有局部可观测马尔可夫性,通信协商和信度分配是多机器人系统的核心,其学习的难度与计算复杂度都远高于单机器人的强化学习,如何考虑多机器人间的知识和共享信息,保证在一定学习速率下加强相互间协作成为技术难题。

传统的 Q 学习算法在初始化过程中将 Q 值设为均等值或随机值,即在无先验知识的环境下进行学习。同时,学习率一般初始化后不会动态调整,学习过程的决策空间大、学习速度慢、学习效果不确定,同时由于每个机器人的感知能力和范围有限,如何克服状态局部感知与不确定性,减少机器人所用的勘探时间也成为一大难题[26-27]。现有技术改进如有 MaxQ 算法、分层强化学习算法、压力学习、关系强化学习、逆强化学习等[28-31]方法来加强和评估多智能体的学习过程,调整强化学习算法相应的学习策略和回报,取得一定成效。Bogert 系统地提出一种最大熵逆强化学习方法,试图建立对移动机器人的偏好、行为的关系,并能很好地预测它们的未来位置,从而提高多个交互机器人在环境被遮挡时的学习能力。Li 等[32]研究了多机器人对具有共同期望轨迹的目标协调控制的问题,采用强化学习考虑了轨迹跟踪和控制输入最小化来处理不同机器人之间分歧问题。但随着多机器人数量的增加,机器人间交互关系愈加复杂,其学习过程中通常存在如下几个问题:1)机器人由于个体性能局限,往往具有邻域结构等特征,在局部范围内进行信息交互,在学习的试错和迭代过程中,消耗了大量的计算资源;2)机器人个体间需要协同合作,其信息交互需占用较大的通信带宽;3)学习过程中各智能体间的联合状态和联合动作的感知和相互影响,使得学习策略随状态、动作维数过高,从而导致结构信度分配难、均衡点选择困难等问题[33]。

本文针对现有多机器人路径规划协同控制过程中存在的问题。提出了一种利用机器人间信息作为启发因子的局部环境感知交互式强化学习方法。首先根据多机器人对范围内目标检测时所产生响应的时效和强度不同,分别定义了目标检测的个体触发响应函数和协同触发响应函数。其中个体触发响应函数目的在于快速确定个体机器人的搜索范围,增强对目标信号的感知,提高个体的探索能力。协同触发响应函数反映了邻域内多个机器人通过交换即时状态、学习策略等信息来改变和提高个体的学习效率。将响应函数引入到分布式马尔可夫模型,设计了个体-协同触发强化学习算法。在多机器人系统的 Q 学习过程中,本文引入双学习率策略改进 Q 学习算法,当多个机器人间感应并触发协同响应函数时,原机器人的检测信号强度及信息作为先验知识协同到追随的个体,后者将选择新的学习率 β 对目标进行学习,以改变原来学习强度和行为特征,提高个体间的协同效率。通过响应触发后计算联合策略,减少机器人间通讯量和计算资源。最后,将该算法应用于多机器人联合搜索路径规划的行为决策,分别对传统 Q 学习方法及文献[34]的改进强化学习进行了对比实验分析,结合实验室环境智能轮式移动小车开展工程实验设计和仿真结果,论证了本文算法的优缺点及适用范围。

1 目标问题分析及建模

相对单智能体强化学习,多智能体强化学习更适合

多机器人协作的复杂问题,MDPs 是该类学习方法的数学模型基础。该类多机器人强化学习结构如图 1 所示。

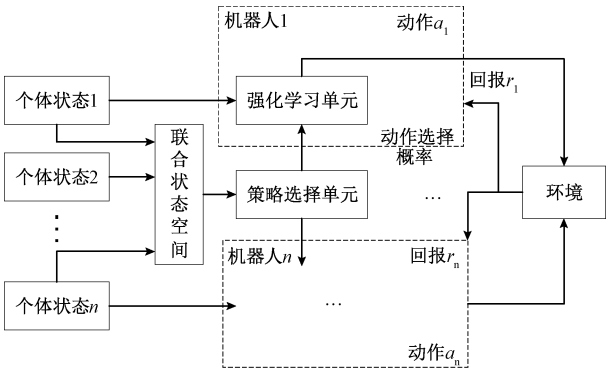


图 1 多机器人强化学习结构

Fig.1 Multi-robot reinforcement learning structure

将多机器人强化学习过程一般化为马尔可夫随机策略,可以用多元组 M: $< S, A^i, p^i, g^i >$ 表示,$i = 1,2,\cdots,n$。其中,n 为机器人个数;S 代表环境状态集合,$s_t^i \in S$ 表示第 i 个机器人在 t 时刻所处的状态; A^i 为机器人个体可选择的动作集合,$a_t^i \in A$ 表示第 i 个机器人在 t 时刻的动作。多机器人联合动作集可表示为 $\boldsymbol{A} = A^1 \times \cdots \times A^n$,$p^i$: $S \times \boldsymbol{A} \times S \rightarrow [0,1]$ 为状态转移概率函数;g^i: $S \times \boldsymbol{A} \times S \rightarrow \boldsymbol{R}$ 为回报函数,表示机器人个体 i 在状态 s_t^i 执行动作 a_t^i 到动作 s_{t+1}^i 得到的立即回报。设策略 π: $S^i \rightarrow A^i$ 为联合状态 $\vec{s} = (s^1, \cdots, s^n)$ 到联合动作空间 $\vec{a} = (a^1, \cdots, a^n)$ 的一个映射,以使行为从环境中获得的累积回报值如式(1)所示。

$$R_t^i(\pi) = r_t^i + \gamma r_{t+1}^i + \gamma^2 r_{t+2}^i + \cdots + \gamma^j r_{t+j}^i = \sum_{j=0}^{T} \gamma^j r_{t+j}^i \tag{1}$$

式中: $\gamma \in [0,1]$, 代表折扣因子。

其在策略 π 下迭代学习规则如式(2)所示。

$$Q_{t+1}^i(\vec{s}_{t+1}, \vec{a}_{t+1}) = Q_t^i(\vec{s}_t, \vec{a}_t) + \alpha[r_{t+1}^i + \gamma \max_{a \in A^i} Q_{t+1}^i(\vec{s}_{t+1}, \vec{a}_{t+1}) - Q_t^i(\vec{s}_t, \vec{a}_t)] \tag{2}$$

在复杂环境中每个机器人的感知能力有限,获得的信息往往带有局部性和不确定性。局部信息可观测的多机器人系统构成了一个分布式马尔可夫模型。学习过程中各智能体间的联合状态和联合动作的感知和相互影响,使得学习策略随状态、动作维数呈指数增长,容易导致信度分配难、维数灾害、均衡点选择困难等新问题。因此,大部分基于值函数的多机器人强化学习算法或基于联合状态-动作的多智能体学习算法都面临学习空间大与稀疏信号之间的矛盾,学习效果不佳。

2 多机器人局部感知触发机制

多机器人在协作工作过程中,通过自身配备的传感元件获取信息而改变自身状态来获得对环境的适应性。单个机器人如果能获取并响应其他机器人的知识,在局部环境中对强化学习回报进行优化,将有助于在保证学习速度的前提下加强它们之间的协作。在多智能体系统的强化学习中,个体间通过交换即时状态、学习策略等信息可提高个体的学习效率。因此,设计多机器人感知交互结构如图 2 所示。

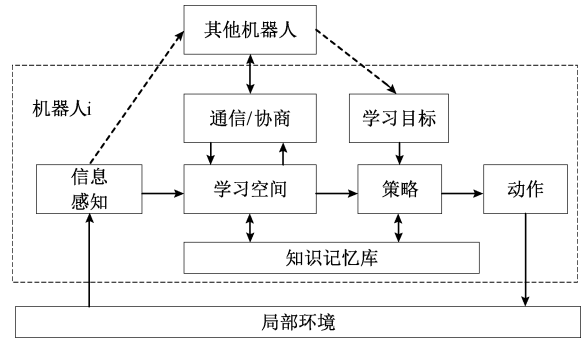


图 2 多机器人感知交互学习

Fig.2 Multi-robot perceptual interactive learning

2.1 个体-协同感知触发函数

本文以多机器人协同搜索为例,通过各机器人自检测目标信号强度,设计自触发和协同触发函数,来表述机器人与最强目标信号的位置之间的距离因素。设 m_k 为同时检测到目标信号 k 的机器人数量;d_{ik} 表示为第 i 机器人检测到信号 k 时的距离;$T_{ik}(d)$ 以距离为变量的函数,定义为式(3)表示第 i 机器人检测到目标信号 k 的强度;θ_{ik} 为第 i 个机器人对检测信号的响应阈值。定义机器人 i 个体感知触发响应函数如式(4)所示。

$$T_{ik}(d) = \begin{cases} 0, & d_{ik} > r \\ P_k / d_{ik}^2 - \vartheta, & d_{ik} \leqslant r \end{cases} \tag{3}$$

$$f(T_{ik}) = \frac{T_{ik}}{(T_{ik}^2 + m_k \theta_{ik}^2 + \eta d_{ik}^2)} \tag{4}$$

式中: P_k 是目标信号发射的能量; r 是传感器检测半径; $d_{ik} \leqslant r$ 时则信号强度与距离平方成反比,若 $d_{ik} > r$,信号强度为 0,代表未检测到目标;η 为距离调节参数;ϑ 是正态分布的白噪声样本。单位时间 t 内,若有邻近机器人 j 检测到目标信号 k 时,以 T_{jk}^* 作为协同触发响应协作参与目标 k 搜索,对于函数式(5)所示。

$$f(T_{jk}^*) = \frac{T_{jk}^*}{(T_{jk}^{*2} + m_k \theta_{jk}^2 + \eta d_{jk}^{*2})} \tag{5}$$

根据个体响应触发函数值来确定单个体机器人的搜索状态变化,通过协同响应触发函数值来调节群体机器人联合状态变化。距离目标信号强度越强,则强化学习回报值越大,反之则越小。

2.2　多机器人的探索-利用机制

由于多机器人感知范围及通讯能力限制,本文通过调节式(4)、(5)的距离响应参数 η 来控制机器人对目标及相互间的感知强度。

在初次检测到目标时,由于对距离及信号响应信息有限,需要快速拓展多机器人的搜索范围,增强对目标信号的感知,提高个体的探索能力,并给予较大的奖励回报。在新目标信号被检测到后,根据式(4)计算各机器人对目标的个体感知触发响应概率,单个体机器人根据设定的阈值决定是否参与目标搜索。当前机器人如果未响应触发和选择搜索,则根据式(5)检测领域机器人的协同触发响应信号,以确定是否参与其他机器的协同搜索。当确定参与目标信号搜索后,多机器人在局部空间对目标进行定位搜索,需要提高个体的利用能力,并根据检测距离调整奖励回报。

图 3 所示描述了多机器人协同参与目标搜索探索-利用过程。假设多机器人在配置和能力上是同构的,同时单位时间内机器人同时只检测到一个目标信号 j。O_j 为信号 j 的位置;m_i 为参与目标搜索的机器人个体,$i=1,2,\cdots,5$;m_1 为当前位置检测到 O_j 信号最强的机器人,代表该机器人将以 $f(T_{1j})$ 概率值进行目标搜索;d_{1i} 代表 m_1 与领域机器人 i 之间的距离。当 m_2 和 m_3 此时也捕获 O_j 信号,根据式(5)分别计算协同响应触发函数,由于 $d_{12}^* < d_{13}^*$,$T_{2j}^* > T_{3j}^*$,从而 $f(T_{2j}^*) > f(T_{3j}^*)$,表明 m_2 获得的信号响应概率大于 m_3。当某一时刻,某个机器人检测到目标信号,通过个体触发响应概率控制机器人的搜索范围,并启发式地加快邻域内其他机器人的检测。当机器人感知多个信号源时,则根据计算协同触发响应概率来判断是否参与协同搜索。在个体-协同触发过程,机器人之间不需每一时刻进行通信,因此减少通信消耗。

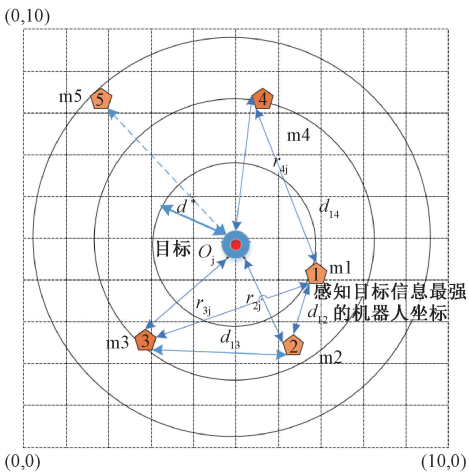


图 3　多机器人目标感知及响应触发

Fig.3　Multi-robot target perception and response triggering

3　双学习率强化学习方法

在第 2 节中介绍的无模型 Q 学习算法中,智能体在每个状态变化过程中都对 Q 值表进行迭代计算,当在多机器人系统及状态空间较大时,其计算过程将十分复杂。本文根据第 3 节设计的个体-协同触发响应函数,通过机器人是否检测到邻域机器人的协同信号,来动态调整个体的学习率,从而引入双学习率 α 与 $\beta(0<\beta<\alpha<1)$ 来对不同过程机器人进行设置,从而改进 Q 强化学习的策略更新。

1)在 t 时刻时,单个机器人个体通过对环境的观测结果计算本身个体-协同触发响应,定义一次触发过程。个体触发的对象是单个机器人。这时,统一对这个时期内的机器人采用学习率 α,采用较大一些的学习率以拓展机器人的状态更新,扩大搜索空间范围。

2)在某机器人检测到邻域机器人的协同响应触发时,是局部环境参与协同工作的多机器人团队相互信号检测并协商通信的结果,其响应触发后的行动是计算联合策略,目的在于减少计算资源消耗。这时,对该检测到协同触发响应信号的机器人采用学习率 β。该学习率小于 α,机器人的学习规则更新发生了变化,加上协同触发响应函数的干预,增强该机器人在局部空间的搜索能力。

因此,重新定义五元组模型:

多元组 M:$<S,A^i,p^i,g^i,f^i>$,其中 f^i 表示响应个体响应触发函数,如式(4)所示,表明个体响应触发的情况下,开始对 Q 值迭代计算。定义在 s_t 时刻的 Q 函数表达式如(6)所示。

$$Q_{t+1}^i(\vec{s}_t,\vec{a}_t,f) = r_t \cdot \max_{a_t}\{Q_{t+1}^i(\vec{s}_t,\vec{a}_t,f) \mid \vec{a}_t \in A^i\} \tag{6}$$

其更新规则为如式(7)所示。

$$Q_{t+1}^i(\vec{s}_{t+1},\vec{a}_{t+1}) = (1-\alpha)Q_{t-l}^i(\vec{s}_t,\vec{a}_t,f) + \alpha[r_{t+1}^i + \gamma \max_{a\in A^i} Q_{t+1}^i(\vec{s}_{t+1},\vec{a}_{t+1},f)] \tag{7}$$

式中:l 表示个体响应触发和当前时刻的差值。当机器人个体没有响应触发时,将不通过式(7)更新 Q 值,而直接选择当前状态 Q 值。当邻近机器人 j 感应到协同触发函数响应时,这些个体选择学习率 β 对目标进行学习,其中 $\beta < \alpha$。

$$Q_{t+1}^i(\vec{s}_{t+1},\vec{a}_{t+1}) = (1-\beta)Q_{t-l}^i(\vec{s}_t,\vec{a}_t,f^*) + \beta[r_{t+1}^i + \gamma \max_{a\in A^i} Q_{t+1}^i(\vec{s}_{t+1},\vec{a}_{t+1},f^*)] \tag{8}$$

触发响应强化学习算法流程如图 4 所示。

4　仿真实验

4.1　单机器人算法执行效率实验

以机器人目标搜索路径规划为例,采用栅格法建立一

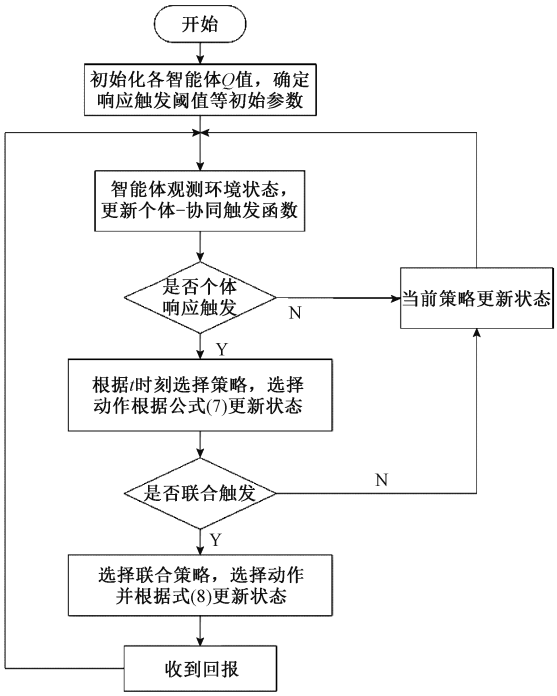


图 4 触发响应强化学习算法流程

Fig.4 Flow chart of trigger response reinforcement learning algorithm

个输入由 0 和 1 组成的 $W\times W$ 的矩阵实验环境。待搜索目标 O_j 的栅格地图模型如图 5 所示，建模空间为一个 200×200 单位二维区域 RS，区域中分布了大小不同的伪目标(障碍物)。其中用 0(空白)表示此处可以通过的，1(阴影)表示此处为障碍物，多机器人以等步长 R 进行移动。实验起始坐标点在第一个栅格 Ori (0,0)，待搜索目标信号坐标为 O_j(160,180)。在实验相同条件下情况下，开展机器人协同目标搜索仿真实验。分别将本文算法与传统 Q 学习算法、改进强化学习算法系列[34]运行比较。

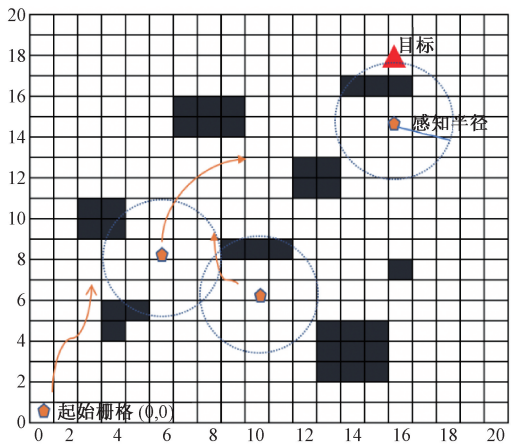


图 5 工作空间栅格

Fig.5 Workspace grid schematic diagram

为便于实验比较，如文献[34]方法初始化地图及实验参数，每一个栅格作为一个状态，每个状态下具有上、下、左、右 4 个动作，将机器人每一步步长 R 设定为一个标准栅格距离，即在 Q 学习策略的下一状态栅格是与当前状态相邻的栅格。设定机器人数量为 1，每次搜索过程的迭代次数为 200。计算机器人平均路径长度 $\bar{S}$(该长度按单位栅格计算)，机器人的移动单位距离的平均时间 - 时步比 $\bar{T}$ 以及平均收敛代数 $\bar{M}$ 用于算法性能评估。考虑到环境误差影响，仿真 20 次取平均值结果如表 1 所示。

表 1 算法执行效率仿真结果对比

Table 1 Comparison of simulation results of algorithm execution efficiency

优化方法	N	$\bar{s}$(栅格)	$\bar{T}$/s	$\bar{M}$
本文算法	1	32.36	1.732	245
QL1 算法[34]	1	79.78	2.714	629
QL2 算法[34]	1	32.7	1.652	274
QL4 算法[34]	1	38.0	1.985	559

由表 1 可知，本文算法较参考文献中标准 Q 学习算法 QL1 相比，算法的平均搜索路径距离短，优化时间短，搜索所需的单位时步比较小，在同等实验条件下，平均性能分别提高了 53%、47% 及 63%。对比文献中的 QL2 算法增加了动作步长和动作数，增加状态集的数量会使算法的收敛速度减慢，但是步长的增加使得 QL2 算法在平均移动路径和单位时步比得到一定优化，但平均移动距离及收敛代数略弱于本文算法。对比文献中的 QL4 算法在 Q 学习过程初始化的过程中引入引力势场初始化 Q 值，减低了算法的平均收敛时间，但平均移动距离增大。

以上实验由于选择的是单个机器人进行的算法测试，仅仅依托个体触发响应函数能有效提高个体在早期的探索能力，并给以较大的奖励回报提高算法收敛性。但协同触发响应机制没有发生作用，在系统整体效率上仅比传统 Q 学习方法取得一定优势。

4.2 多机器人协同搜索实验分析

本文算法在强化学习过程中增设了学习率β($\beta<\alpha$)，同时通过触发响应机制进行动态调整学习率，改变了 Q 表的初始化过程中将其设为均等或完全随机带来的问题。实验 2 采用不同学习率的情况下，测试了算法收敛性并对多机器人强化学习协同效率进行了对比分析。实验过程的初始参数选择如表 2 所示。

表 2　实验参数

Table 2　Experiment parameters

参数	定义	值
RS	建模空间	200×200 dm
Ori	起始坐标点	(0,0) dm
R	单位步长	1 dm
r	传感器检测半径	10 dm
Target	目标位置	(160,180) dm
θ	响应阈值	0.1
η	距离调节因子	0.006~1
N	机器人数量	2~10 个
α	学习因子	0.6~0.8
β(本文算法专有)	学习因子	0.3~0.5

图 6 所示描述了两种算法在不同学习率情况下对算法收敛性能的影响。

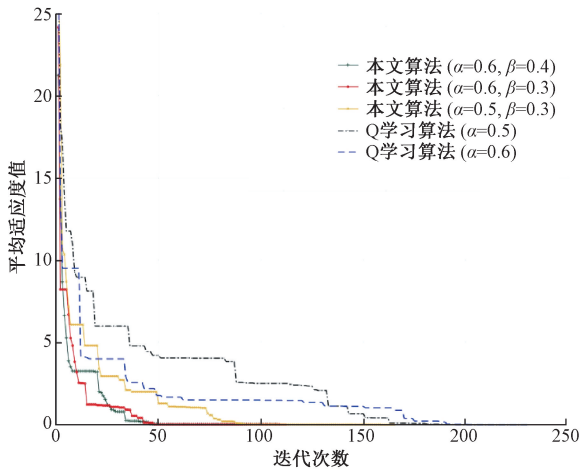


图 6　平均适应值进化曲线

Fig.6　Mean fitness value evolution curve

如图 6 所示,设计了 3 组不同学习率组合与两组固定学习率 Q 学习算法比较。本文算法在平均适应度上要优于固定学习率 Q 学习算法。在不同的学习率控制下,本文算法在后期收敛速率快而且收敛精度高,平均在 100 代内学习到最优值。在 $\alpha=0.6$,$\beta=0.3$ 情况下,本文算法到 46 代即收敛于最优,距离误差为 0.81%。传统 Q 学习算法接近 170 代才收敛,并且在收敛时末期有振荡出现,表示机器人是在目标附近徘徊。

进一步实验验证机器人数量选择对本文强化学习效率将产生影响,当 N 在 5~40 间取值变化时,算法的执行效率比较,如表 3 所示。

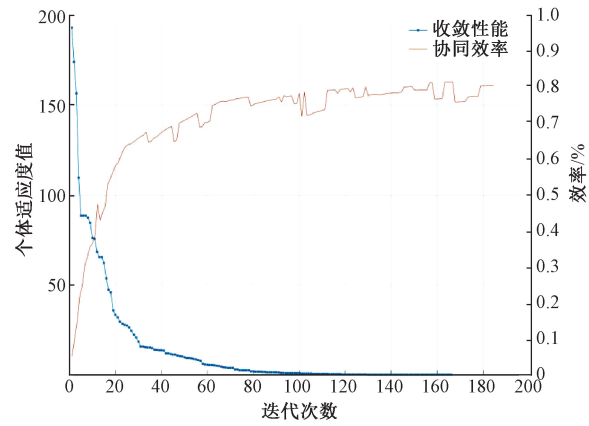
表 3　机器人数量 N 取不同值时本文算法仿真结果

Table 3　Simulation results of the proposed algorithm when the number of robots N takes different values

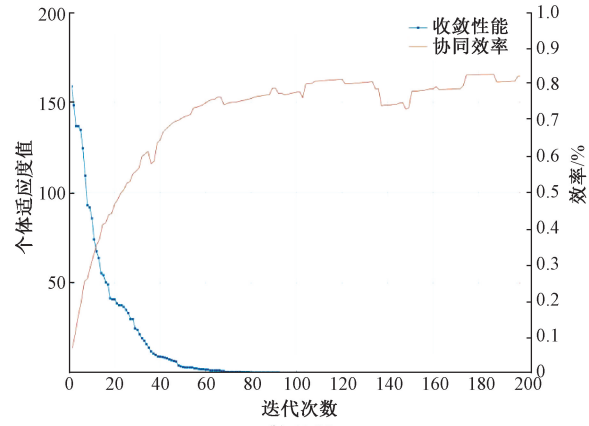
N	仿真次数	迭代代数	$\bar{s}$	$\overline{T}$/s	$\overline{M}$
5	500	200	54.171 1	2.022 0	92
10	500	200	51.471 4	1.188 0	83
15	500	200	45.937 0	1.141 3	75
20	500	200	42.290 2	1.085 0	65
25	500	200	43.874 3	1.154 4	69
30	500	200	45.733 6	1.188 5	83
40	500	200	53.108 9	1.805 1	138

如图 7 所示,对 N 分别取 5、20、40 时,分析比较了算法收敛性及机器人协同效率。

1) 当 $N=5$ 时,多机器人平均搜索路径长度 $\bar{S}$ 最大,为 54.17 个单位,时步比 $\overline{T}$ 以及平均收敛代数 $\overline{M}$ 也相对较大。当 N 取值逐步增大,随着加入联合搜索的机器人节点增加,协同搜索效果得到增强。平均移动距离和时步比都相应降低。



(a) N=5



(b) N=20

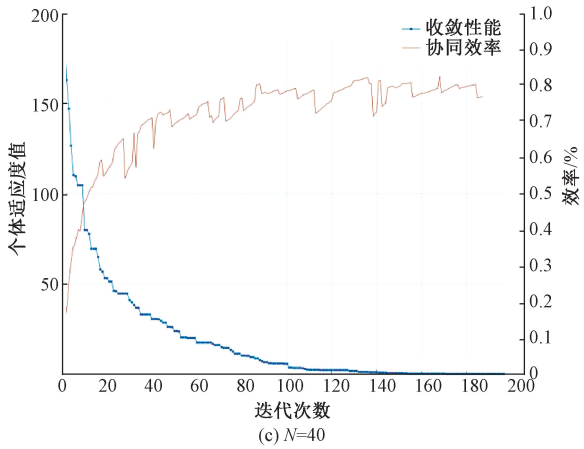


图 7 机器人数量对协同效率及系统性能的影响

Fig.7 Influence of the number of the robots on cooperative efficiency and system performance

2）当 $N=20$ 时，本文算法所得结果最优。单位时步比为 1.085 0。但随着 N 继续增大，在有限空间内，机器人相互耦合增强，协同触发函数交叉响应率提高，干扰增强，导致有效通信效率降低，搜索效率反而降低。同时当进一步增大 N 时，$\bar{S}$ 则随着 N 增大一直增加，这说明随着 N 增大，时变特征群增多，多机器人所分布区间面积增大，从而导致群体间信息交互复杂，收敛路径长度增加，系统易出现不稳定。

考虑到栅格图分辨率及环境设置对仿真结果影响较大。可以通过协同触发响应函数中距离调节变量 η 和响应阈值 θ_{ik} 来控制机器人相互间的感知强度及敏感度，从而改变响应触发的概率和协同效率。采用表 2 所示参数初始化系统，其中机器人个数 $N=6$，η 分别取 0.001，0.005，0.008，0.01，0.05，0.1 时，用机器人的移动时间步长比和平均移动距离（单位栅格）来衡量搜索效率和能耗。$\theta_{ik}=0.1$ 及 0.2 的情况下，分别运行 20 次，仿真结果如表 4 所示。

表 4 距离调节因子 η 对算法影响分析

Table 4 Influence analysis of distance adjustment factor η on the algorithm

η	响应阈值 $\theta_{ik}=0.1$		响应阈值 $\theta_{ik}=0.2$	
	$\bar{T}$/s	$\bar{s}$	$\bar{T}$/s	$\bar{s}$
0.001	2.333	55.655 0	2.238	55.943
0.005	2.022	54.171 1	2.043	55.673
0.008	1.982	53.322	2.036	55.345
0.01	2.032	53.430	2.045	55.984
0.05	2.472	57.453	2.564	57.854
0.1	2.543	58.321	2.642	58.002

根据表 4 结果，响应阈值 θ_{ik} 对算法执行效率影响较小，但当阈值设置过大时，将会导致机器人忽略掉协同响应型号，因此一般 θ_{ik} 取值小于 0.6。但距离变量调节因子 η 取值在 0.001 ~ 0.008 之间变化时，机器人完成目标搜索任务时所需的时步比呈略微下降，但变化幅度很小。当 $\eta=0.008$ 时，所需的移动时步比和平均移动距离都能达到最小值，而当 η 取值进一步增大时，移动时步比和平均移动距离又呈上升趋势。

4.3 移动轮式搜救机器人物理实验

目前灾害应急搜救群机器人系统是一种高度集成的智能化系统，配备有对辅助环境根感知设备，能将图像、压力、温度和障碍等信息进行综合分析并进行协同处理能力，从而完成对指定（人或物）进行快速搜救任务。本文实验采用基于 Raspberry pi3 为主控平台的轮式智能机器人作为物理实验对象，组成目标搜救群机器人系统。每个机器人系统配置有 4 个基础传感器，包括一组 360° 检测的碰撞传感器、一组光敏传感器、一组声音传感、红外传感和检测轮子转动角速度的光电编码器。分别完成对模拟环境的声光、避障和稳定的复杂环境监测，并进行实时信息分析。机器人配备 256 M Flash 存储，支持 MicroPython 在线编程调试。

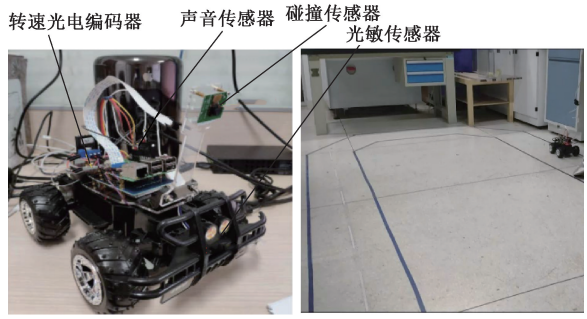


图 8 物理仿真实验环境

Fig.8 Physical simulation experiment environment

实验在范围为 800 cm×800 cm 光滑平整的实验场地中进行，由于空间局限，实验过程没有设定障碍物。在坐标 $(X,Y)=(600,600)$ cm 设置一个稳定的声源作为目标信号，发声器以稳定时间间隔发出声音，忽略实验环境放射产生的回音及噪音干扰。其具体实验参数如表 5 所示。

将智能小车机器人随机初始分布在一定半径（280，280）cm 的起始区域，随机向目标信号搜索。为便于分析个体-协同触发响应机制对多机器人行为控制效果，采用移动终端控制系统通过 ZigBee 协议无线通信模块实时采集获取机器人的传感器数据和机器人的状态。机器人的坐标分配、概率响应强度计算及信息交互由平台监控系统来完成，如图 9 所示。控制系统通过 ZigBee 协议无

表 5　Raspberry Pi 轮式机器人实验参数

Table 5　Experiment parameters of Raspberry Pi wheeled trolley

类别	参数描述
轮式机器人数量/个	10
信号属性	声源
搜索目标坐标/cm	$(X,Y)=(600, 600)$
光源	红外点阵光源 3 个
实验范围/cm	800×800
初始位置/cm	$(X,Y)=(0,0)-(40,40)$

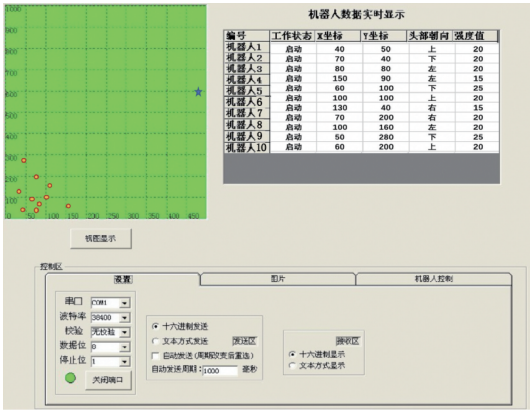


图 9　机器人系统仿真监控平台

Fig.9　Robot system simulation monitoring platform

线通信模块实时采集获取机器人的传感器数据和机器人的状态,并以不同颜色的圆点代表不同的成员机器人,其运动轨迹在 PC 端的监控系统界面进行显示。

由图 10(a)(本文算法)及图 10(b)(Q 学习算法)所示行为轨迹仿真实验可知,在目标搜索初期,机器人初始化分布所在区域距离目标较远(传感器感知半径 r 无法覆盖到目标信号),任何一个机器人都无法即刻感知目标,机器人扩大搜索范围开展随机搜索。在这个阶段,当某一个机器人检测到一个邻近个体或多个个体响应触发信号时,机器人会较快响应并可能调整原搜索方向,以免搜索行为过于发散。同时由于触发了协同响应函数,当前机器人的强化学习将改变原始的学习率,采用新的学习率进目标进行学习并更新当前状态 Q 值,减少状态空间计算复杂度。当其中有一个机器人检测到目标时(以传感器半径 r 圆可覆盖目标信号),通过与其时变特征群中的机器人进行信息交互,触发协同响应概率函数加速其他机器人对搜索目标的感知,促使越来越多的机器人减少前期探索时间,提高了收敛速率和趋近目标的精度。多机器人从初始起点随机向目标信号搜索,目标搜索初期时其竖切面较宽,在趋向目标时越来越窄,表明算法在学习起始阶段具有较强的全局探索能力,而在学习的后期阶段则具备较强的局部开发利用能力。

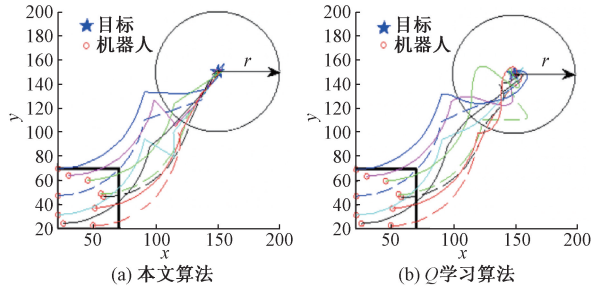


图 10　协同搜索行为轨迹

Fig.10　Cooperative search behavior track

5　结　　论

本文提出了一种局部环境感知交互的分布式强化学习方法,应用于多机器人联合目标搜索的行为控制策略。采用栅格法对机器人工作环境进行建模,基于个体的有限感知能力和局部的交互机制设计了响应概率函数,通过通信获取相互间感知目标信号的强度,解决群机器人任务分配与信息共享难题。提出了一种个体-协同触发强化学习方法,并将其引入分布式马尔可夫模型,运用概率预测函数降低强化学习的维数以加快收敛速度。引入双学习率策略改进 Q 学习算法,通过触发后计算联合策略,减少机器人间通讯量和计算资源。最后,将该算法应用于多机器人的联合搜索行为决策并进行了对比实验分析,显示本文算法较传统 Q 学习算法平均搜索路径距离短,优化所用时间短,收敛代数较小,对于提高多机器人协作能力具有较好参考价值。

参考文献

[1] 张大伟,孟森森,邓计才.多移动微小型机器人编队控制与协作避碰研究[J].仪器仪表学报,2017,38(3):578-585.
ZHANG D W, MENG S S, DENG J C.Formation control and cooperative collision avoidance for multiple mobile miniature robots [J]. Chinese Journal of Scientific Instrument, 2017,38(3): 578-585.

[2] 段勇,徐心和. 基于多智能体强化学习的多机器人协作策略研究[J].系统工程理论与实践,2014,34(5):1305-1320.
DUAN Y, XU X H. Research on multi-robot collaboration strategy based on multi-agent reinforcement learning [J]. System Engineering Theory and Practice, 2014, 34(5): 1305-1320.

[3] CHEN C L, DONG D Y, LI H X. Fidelity based

probabilistic Q-learning for control of quantum systems[J]. IEEE Transactions on Neural Networks and Learning Systems,2014,25(5): 920-933.

[4] 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): 1601-1610.
WU J, XU X, WANG J, et al. Recent advances of reinforcement learning in multi-robot systems: A survey[J]. Control and Decision, 2011,26(11):1601-1610.

[5] 刘涛,王淑灵,詹乃军. 多机器人路径规划的安全性验证[J]. 软件学报,2017,28(5):1118-1127
LIU T, WANG SH L, ZHANG N J. Safety verification of trajectory planning for multiple robots[J]. Journal of Software. 2017,28(5):1118-1127.

[6] CAI Y F, YANG S X, XU X. A hierarchical reinforcement learning-based approach to multi-robot cooperation for target searching in unknown environments[J]. Control and Intelligent Systems, 2013,41(4): 218-30.

[7] 徐雪松,杨胜杰,陈荣元. 复杂环境移动群机器人最优路径规划方法[J]. 电子测量与仪器报,2016,30(2): 274-282.
XU X S, YANG SH J, CHEN R Y.Dynamic differential evolution algorithm for swarm robots search path planning[J]. Journal of Electronic Measurement and Instrumentation,2016,30(2): 274-282.

[8] 易国,毛建旭,王耀南,等.多移动机器人运动目标环绕与避障控制[J].仪器仪表学报,2018,39(2):11-20.
YI G, MAO J X, WANG Y N,et al. Circumnavigation of a moving target and obstacle avoidance with multiple mobile robots [J], Chinese Journal of Scientific Instrument,2018, 39(2):11-20.

[9] 张大伟,孟森森,邓计才.多移动微小型机器人编队控制与协作避碰研究[J].仪器仪表学报,2017,38(3): 578-585.
ZHANG D W, MENG S S, DENG J C. Formation control and cooperative collision avoidance for multiple mobile miniature robots [J]. Chinese Journal of Scientific Instrument,2017,38(3):578-585

[10] BOGERT K, DOSHI P. Multi-robot inverse reinforcement learning under occlusion with estimation of state transitions[J]. Artificial Intelligence,2018. 263:46-73.

[11] 陈彦杰,王耀南,谭建豪,等.局部环境增量采样的服务机器人路径规划[J]. 仪器仪表学报,2017,38(5): 1093-1100.
CHEN Y J, WANG Y N, TAN J H, et al. Recent advances of reinforcement learning in multi-robot systems: A survey [J]. Chinese Journal of Scientific Instrument, 2017,38(5): 1093-1100.

[12] 刘志荣,姜树海.基于强化学习的移动机器人路径规划研究综述[J].制造业自动化,2019,41(3):90-92.
LIU ZH R, JIANG SH H. Review of mobile robot path planning based on reinforcement learning [J]. Manufacturing Automation, 2019,41(3):90-92.

[13] CUI Y D, TAKAMITSU M, KENJI S. Kernel dynamic policy programming: Applicable reinforcement learning to robot systems with high dimensional states[J]. Neural Networks, 2017,94(10): 13-23.

[14] 马磊,张文旭,戴朝华. 多机器人系统强化学习研究综述[J]. 西南交通大学学报,2014,49(6):1032-1044.
MA L, ZHANG W X, DAI CH H. A review of developments in reinforcement learning for multi-robot systems[J]. Journal of Southwest Jiaotong University, 2014,49(6):1032-1044.

[15] SHAO J, DU L J, LIN H X. Multi-robot reinforcement learning based on LCS and LS-SVM [J]. Artificial Intelligence and Robotics Research, 2013(2):24-28.

[16] GU S, LILLICRAP T, SUTSKEVER I. Continuous deep q-learning with model-based acceleration [C]. Proceedings of the 32nd International Conference on Machine Learning. New York, USA, 2016: 2829-2838.

[17] LAKSHMANAN A K, MOHAN R E, RAMALINGAM B, et al. Complete coverage path planning using reinforcement learning for tetromino based cleaning and maintenance robot [J]. Automation in Construction, 2020,112(5):456-478.

[18] 张浩杰,苏治宝,苏波.基于深度 Q 网络学习的机器人端到端控制方法[J].仪器仪表学报,2018,39(10): 36-43.
ZHANG H J, SU Z B, SU B. End to end control method of robot based on deep Q learning network[J]. Chinese Journal of Scientific Instrument, 2018,39(10):36-43

[19] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518(7540): 529-533.

[20] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning. Computer Science, 2016, 8(6): 178-187.

[21] 朱美强,李明,程玉虎,等.基于拉普拉斯特征映射的启发式 Q 学习[J]. 控制与决策,2014,29(3): 425-430.
ZHU M Q, LI M, CHENG Y H, et al. Heuristically accelerated Q-learning algorithm based on laplacianeigenmap[J]. Control and Decision, 2014, 29(3): 425-430.

[22] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and

tree search[J]. Nature, 2016,529(7587): 484-489.

[23] FANG M, GROENF C A, LI H. Collaborative multiagent reinforcement learning based on a novel coordination tree frame with dynamic partition [J]. Engineering Applications of Artificial Intelligence, 2014, 27(7): 191-198.

[24] TAMPUU A, MATIISEN T, KODELJA D, et al. Multiagent cooperation and competition with deep reinforcement learning [J]. PLoS ONE, 12(4): e0172395.

[25] ADOLFO P, WEN Y, ALBERTO S. Position/force control of robot manipulators using reinforcement lear-ning[J]. The Industrial Robot, 2019, 46(2): 267-280.

[26] 徐继宁,曾杰. 基于深度强化算法的机器人动态目标点跟随研究[J].计算机科学, 2019,46(S2):94-97.
XU J N, ZENG J. Dynamic target following based on reinforcement learning of robot-car [J]. Computer Science, 2019, 46(S2): 94-97.

[27] 张浩杰,苏治宝,苏波.基于深度 Q 网络学习的机器人端到端控制方法[J].仪器仪表学报,2018,39(10):36-43.
ZHANG H J, SU ZH B, SU B. End to end control method for mobile robots based on deep Q network[J]. Chinese Journal of Scientific Instrument, 2018, 39(10): 36-43.

[28] KAR S, MOUAJM F, POOR H V. QD-Learning: a collaborative distributed strategy for multi-agent reinforcement learning through consensus plus innovations[J]. IEEE Transactions on Signal Processing, 2013, 61(7): 1848-1862.

[29] STÉPHANE L, BENOIT M, PABLO M. Neural network based reinforcement learning for audio-visual gaze control in human-robot interaction [J]. Pattern Recognition Letters, 2018. DOI: 10.1109/IROS.2018.8594327.

[30] ZHOU X M, BAI T, GAO Y B. Vision-based robot navigation through combining unsupervised learning and hierarchical reinforcement learning [J]. Sensors, 2019, 19(7): 1576.

[31] GUO W, ROBERT B, POLLEY D B. The cholinergic basal forebrain links auditory stimuli with delayed reinforcement to support learning [J]. Neuron, 2019, 103: 1164-1177.

[32] LI Y N, CHEN L, TEE K P, LI Q Q. Reinforcement learning control for coordinated manipulation of multi-robots[J]. Neurocomputing, 2015, 170: 168-175.

[33] 刘全,章鹏,钟珊,等. 连续空间中的一种动作加权行动者评论家算法[J]. 计算机学报, 2017, 40(6): 1252-1264.
LIU Q, ZHANG P, ZHONG SH, et al. An improved actor-critic algorithm in continuous spaces with action weighting [J]. Chinese Journal of Computers, 2017, 40(6): 1252-1264.

[34] 徐晓苏,袁杰. 基于改进强化学习的移动机器人路径规划方法[J].中国惯性技术学报, 2019, 27(3): 314-320.
XU X S, YUAN J. Path planning for mobile robot based on improved reinforcement learning algorithm[J]. Journal of Chinese Inertial Technology, 2019, 27(3): 314-320.

作者简介

徐雪松(通信作者),2001 年、2004 年和 2009 年于湖南大学分别获得学士学位、硕士学位和博士学位。现为湖南工商大学教授,主要研究方向为人工智能,多智能体及复杂系统优化。
E-mail:xuxs@ hutb.edu.cn

Xu Xuesong (Corresponding author) received his B. Sc., M.Sc and Ph. D. degrees all from Hunan University in 2001, 2004 and 2009, respectively. Now, he is a professor in Hunan Technology and Business University. His main research interests include artificial intelligence, multi-agent and complex system optimization.