基于事件驱动的无人机强化学习避障研究

唐博文[a]，王智文*[b]，胡振寰[a]

（广西科技大学 a.电气与信息工程学院；b.计算机科学与通信工程学院，广西 柳州 545006）

摘 要：强化学习方法在避障研究中应用广泛，针对其需要消耗大量的计算资源问题，本文提出一种基于事件驱动的无人机强化学习避障算法.通过在强化学习中加入事件驱动的触发机制，减少无人机的动作决策的同时找到最优路径，既可以保证性能，又可以降低系统的通信频率.实验的仿真结果表明，该算法可以在学习过程中减少对计算资源的消耗，并且完成避障任务的同时可以明显加快收敛速度.

关键词：事件驱动；强化学习；避障；*Q*-learning

中图分类号：TP18 DOI：10.16375/j.cnki.cn45-1395/t.2019.01.015

0 引言

随着无人机在工业、军事及生活等诸多领域的广泛应用[1-2]，人们对其智能化的要求也越来越高，无人机的避障研究越来越被重视.强化学习理论是在观察生物物种的行为学习基础上发展起来的[3]，可以应用在无人机避障算法中.文献[4]使用神经网络（NNs）来进行强化学习，在学习的过程中事件触发机制被设计为估计NN权重的函数.这种设计背后的基本原理是在初始学习期间增加事件，以促进学习.文献[5]提出了一种基于混合学习方案的近似动态规划与在线探索相结合的不确定输入仿射非线性子系统与事件触发状态反馈的分布式控制方案.将在线控制框架中的探索与标识符相结合，以降低总体计算成本，但是在最初的在线学习阶段需要额外的计算.通过调节系统状态和NN权重估计误差来实现局部一致的最终有界结果.强化学习需要强大计算能力作为支撑，如何减少学习中的计算量，是本文研究的重要内容，在此基础上本文提出基于事件驱动的无人机强化学习避障算法.

事件触发机制被设计为估计权重的函数.这种设计背后的基本原理是在初始学习期间增加事件，以促进学习.文献[6]提出了分布式事件触发算法解决一阶多智能体系统的环形编队问题.当执行器信号必须经由公共通信网络频繁交换时，处理器使用率、能耗和通信带宽方面效率低下的挑战会随着这些情况而增加.因此考虑一种替代控制方式，即事件触发控制（ETC），它已经在早期工作中提出并进一步研究[7-8].文献[9]报道了事件触发协议在降低通信频率和控制更新方面的成功应用.文献[10]讨论了在处理包括干扰、时延和网络丢包在内的实际影响时的事件触发机制.文献[11]研究了时间相关的事件触发函数，其中每个代理只需要它自己的确切信息，而不需要其周围环境.文献[12]通过在随机设置中建立一个积分不等式，导出了一个标准用于根据线性矩阵不等式的解来计算合适的事件触发控制器.

目前，把事件驱动和强化学习结合的研究相对较少.因此引入事件触发控制方案可以减少网络负载的数量[8]，信号是否被采样取决于系统状态的某种事件触发条件，而不是时间流逝[13-15].有关事件触发控制的大量结果已经推导出来[16-22].事件触发控制的一个显著特点是，通过连续监测瞬时系统状态或通过在线/离线计算预测某些与状态相关的功能的值，确定下一个采样时刻.文献[23]采用一个评论者网络的*Q*学习框架来近似最优成本和一个零阶保持行为网络来逼近最优控制.本文提出了基于事件驱动的无人机强化学习

收稿日期：2018-08-21

基金项目：国家自然科学基金项目（61462008，61751213，61365009）；广西自然科学基金项目（2014GXNSFAA118368）；广西科技大学创新团队项目（gxkjdxcxtd201504）；柳州市科学研究与技术开发计划项目（2016C050205）；广西科技大学创新项目（GKYC201708）资助.

***通信作者：**王智文，博士，教授，研究方向：机器学习与计算机视觉、移动目标检测与行为识别，E-mail：wzw69@126.com.

避障算法，将基于事件驱动的强化学习运用到无人机避障领域中，在避障的同时优化了算法的资源消耗.

1　强化学习介绍

1.1　强化学习

强化学习（Reinforcement learning）不同于机器学习中的另外两类学习方法（监督学习和非监督学习），其基本思想是借鉴人类学习的过程，让智能体（Agent）通过不断试错来寻找最优策略，即累计回报最大，因此需要设置每种状态及行动对应的回报.

强化学习包含4个主要元素：环境（Environment）、状态（State）、回报（Reward）、行动（Action）.在每个时间点t，智能体都会从可以选择的行动集合A中选择一个行动执行.这个行动集合可以是连续的也可以是离散的.根据图1，在t时刻，st表示无人机当前的状态，α_t表示无人机当前动作，r_t表示当前奖赏值.状态和动作之间存在映射关系，也就是一个状态可以对应一个动作，或者对应不同动作的概率（通常用概率来表示，概率最高的就是最值得执行的动作）.状态与动作的关系其实就是输入与输出的关系，而状态到动作的映射过程被称为策略（Policy）.即强化学习的目标就是找到最优策略使得累计回报和最大.

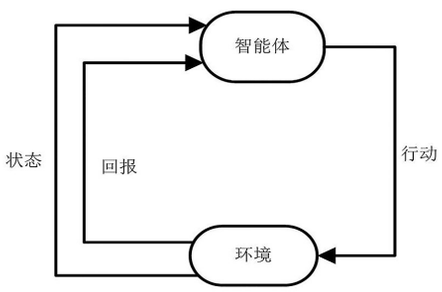

图1　强化学习流程图

Fig.1　Flow chart of reinforcement learning

1.2　Q-learning

Q-table的行和列分别表示状态和行动的值，Q-table的值$Q(s,a)$用来衡量当前状态采取行动到底有多好.在训练的过程中，可以用式（1）贝尔曼方程去更新Q-table.

$$Q(s,a)=r+\gamma(\max(Q(s',a'))) \tag{1}$$

根据贝尔曼方程，学习的最终目的是得到收敛的Q-table，具体算法如下：

Step 1　给定参数γ和回报矩阵R；

Step 2　令Q:=0；

Step 3　对于每个事件：

1）随机选择一个初始的状态s;

2）若未达到目标状态，则执行以下几步：

①在当前状态s的所有可能行为中选取一个行为a;

②利用选定的行为a，得到下一个状态s';

③按照式（1）计算Q（s，a）;

④令s：=s'

基于Q-learning的避障算法通过尝试各种不同的行动来找到最优策略，因此带来了一个很大的问题，那就是算法可能需要遍历所有可能的行动，从而消耗大量的资源.

2　基于事件驱动的强化学习

事件驱动机制已经被证明可以有效减小大规模网络的通信量.根据已有研究成果，事件驱动条件设计主要分为两类：状态相关和状态无关.其主要做法都是通过检测无人机采样前后状态的偏差值大小，判断是否满足事件驱动条件，来决定间歇性的更新控制输入，减小控制器与多智能体系统的通信频率和计算量.综合以上分析，区别于传统的多智能体强化学习算法，在资源有限的情况下，考虑将事件驱动和强化学习相结合，侧重于事件驱动在强化学习策略方面的研究.

基于事件驱动的强化学习过程不同于经典的强化学习，首先需要根据触发函数来判断事件是否被触发，在没有被触发情况下，将直接选用上一个Q值的动作当作当前的Q的动作.

令X是任意集合，并假定B是在X上有界的空间函数$B(X)$，$T{:}B(X)\to B(X)$.设v^*是T的一个固定点，令$T=(T_0,T_1,\cdots)$在v^*近似T，并且对于来自$F_0(v^*)$的初始值，假设在T时F_0是不变的.令$v_0\in F_0(v^*)$，并且定义$V_{t+1}=T_t(V_t,V_t)$.如果存在随机函数$0\leqslant F_t(x)\leqslant 1$和$0\leqslant G_t(x)\leqslant 1$满足式（3）和式（4），则在$B(X)$标准中$V_t$收敛于$v^*$（恒成立）：

1）对于所有的$U_1,U_2\in F_0$和所有的$x\in X$，

$$\left|T_t\left(U_1,v^*\right)(x)-T_t\left(U_2,v^*\right)(x)\right|\leqslant G_t(x)\left|U_1(x)-U_2(x)\right| \tag{2}$$

2）对于所有的$U,V\in F_0$和所有的$x\in X$，

$$\left|T_t\left(U_1,v^*\right)(x)-T_t(U,V)(x)\right|\leqslant F_t(x)(\left|\left|v^*-V\right|\right|+\lambda_t) \tag{3}$$

当$t\to\infty$，$\lambda_t\to\infty$（恒成立）；

3）对于所有的$k>0$和$\prod_{t=k}^{n}G_t(x)$在x中均匀收敛为0；

4）存在$0\leqslant\gamma<1$，使所有的X和足够大的t，满足：

$$F_t(x)\leqslant\gamma(1-G_t(x)) \tag{4}$$

考虑一个具有预期总折扣成本标准的马尔科夫决策过程（MDP），折扣因子$0<\gamma<1$.假设在t时刻有一个四元经验组$\left(x_t,a_t,y_t,c_t\right)$，其中$x_t,y_t\in X$，$a_t\in A$，$c_t\in R$分别是决策者的实际状态和下一个状态，决策者的行为以及在步骤t处收到的随机成本.假设$\left(x_t,a_t,y_t,c_t\right)$始终保持不变.

假设1 考虑有限取样假设（MDP）（X，A，c）.其中$\Pr(y\left|x,a\right.)$是转移概率，$c(x,a,y)$是直接的成本.假设$\{\left(x_t,a_t,y_t,c_t\right)\}$是一个固定的随机过程，令$F_t$作为一个增加的$\alpha$场（历史空间），使序列$\{(x_t,a_t,y_{t-1},c_{t-1},\cdots,x_0)\}$是可测的（$x_0$可以是随机的）.假设以下式子成立：

1）$\Pr(y_t=y\left|x=x_t,a=a_t,F_t\right.)=\Pr(y\left|x,a\right.)$；

2）$E\left[c_t\left|x=x_t,a=a_t,y=y_t,F_t\right.\right]=c\left(x,a,y\right)$且$\mathrm{Var}[\,c_t\left|x_t,a_t,y_t,F_t\right.]$与$t$无关；

3）y_t和c_t独立于历史的F_t.

对应于其中的情况，决策者在真实系统中获得经验，可以设置为$x_{t+1}=y_t$，这与蒙特卡洛模拟形成鲜明对比（$x_{t+1}=y_t$不一定成立）.

$$Q_{t+1}(s,a)=\left(1-\alpha_t(x,a)\right)Q_t(x,a)+\alpha_t(x,a)\,(c_t+\gamma\min_b Q_t(y_t,b)) \tag{5}$$

Q−learning算法由式（5）给出.考虑有限MDP中的Q-learning，其中序列$<x_t,a_t,y_t,c_t>$满足**假设1**.假设学习率序列满足以下条件：

1）$0\leqslant\alpha_t(z,a)$，$\sum_{t=0}^{\infty}\alpha_t(z,a)=\infty$，$\sum_{t=0}^{\infty}\alpha_t^2(z,a)<\infty$，并且均保持一致；

2）如果$(x,a)\neq(x_t,a_t)$，则$\alpha_t\left(z,a\right)=0$.

那么式（5）定义的值收敛到Q^*函数的最优Q值（恒成立）.

3 仿真结果及分析

为了验证本文提出算法和基于强化学习的无人机避障算法的性能，在Windows10操作系统下利用matlab2014a软件进行仿真实验.首先设置一个20×20的迷宫环境（如图2所示），图2对应的Q值如图3所示.假设图2中无人机从绿点出发飞行到红点结束，每个位置飞行都有上下左右4种行动（图2中的箭头所示）可以选择.在探索环境时，如果碰到障碍物，会给予一个很高的惩罚（−50），并且在每次行动过后对迭代的状态进行评分，如果无人机已经飞抵终点，则取消给予惩罚，如果没有到达终点，给予−1的惩罚，以此来不断选取回报最高的动作.在无人机到达终点前重复上述步骤，直到步数确定，可以收敛为止.

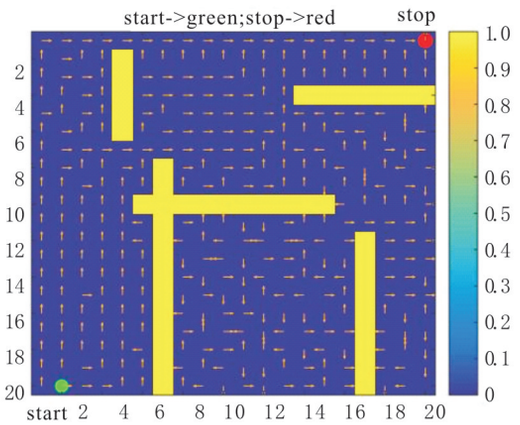

图2 迷宫环境

Fig.2 Labyrinth environment

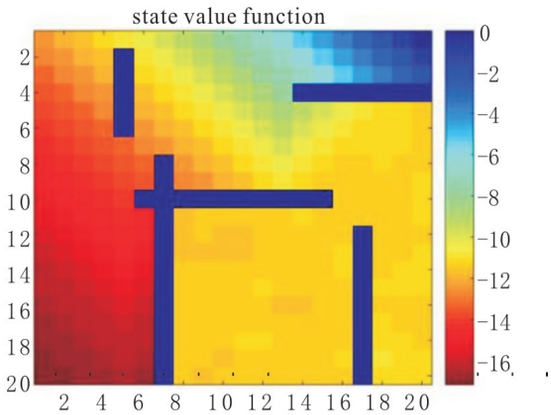

图3 Q 值表图

Fig.3 *Q*-table

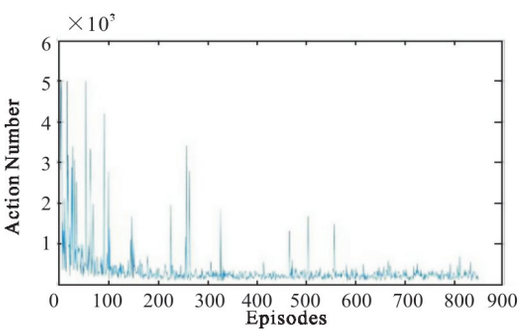

图4 没有事件驱动的迭代次数图

Fig.4 Iteration number map without event driven

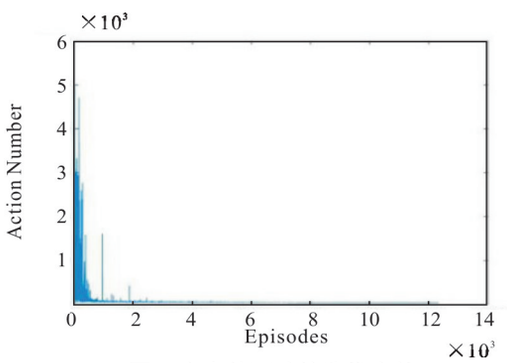

图5 有事件驱动的迭代次数

Fig.5 Iteration number map with event driven

图4和图5分别对应于本文提出算法和基于强化学习的无人机避障算法的迭代次数.对比图4和图5，可以看出，基于事件驱动的无人机强化学习避障算法相比仅包含强化学习的无人机避障算法，收敛速度更快，计算量更少.为了评估3个主要参数对本文提出算法的整体计算量的影响，采用改变其中一个参数并保持另外两个参数不变进行实验，实验结果如表1所示.

从表1可以看出保持两个参数不变，只改变一个参数时，算法的优化率有很大的不同.当学习率为0.3、折扣因子为1、增益系数为0.002时，算法的优化率较好，较原算法减少了198 982次计算，优化率达到66.3%；当学习率低于0.28时，虽然运算次数有很大的减少，但结果会出现不收敛的情况.

表1 测试结果

Tab.1 Results of testing

学习率	折扣因子	增益系数	运算次数（优化前）	运算次数（优化后）	差值	优化率/%
0.30	1.00	0.002	299 981	100 999	198 982	66.3
0.30	0.99	0.002	299 968	103 247	196 721	65.6
0.30	0.98	0.002	299 987	105 197	194 790	64.9
0.30	0.97	0.002	299 981	107 666	192 315	64.1
0.30	1.00	0.003	299 982	116 883	183 099	61.0
0.30	1.00	0.004	299 989	132 323	167 666	55.9
0.30	1.00	0.005	299 979	145 352	154 627	51.5
0.29	1.00	0.002	299 982	95 690	204 292	68.1
0.31	1.00	0.002	299 981	104 084	195 897	65.3
0.28	1.00	0.002	299 958	92 832	207 126	69.0

为了更好地模拟真实环境，通过在地图中设置各种不同的障碍物，如图6的长条迷宫环境，图10的梯形迷宫环境，图14的十字形迷宫环境，然后在这3种不同环境中应用无人机强化学习的避障算法和本文提出的基于事件驱动的无人机强化学习避障算法进行实验.图6、图10、图14对应的实验结果

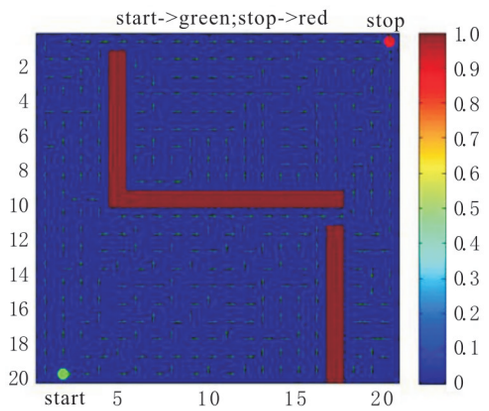

图 6 长条迷宫环境图

Fig.6 Long-shaped labyrinth environment

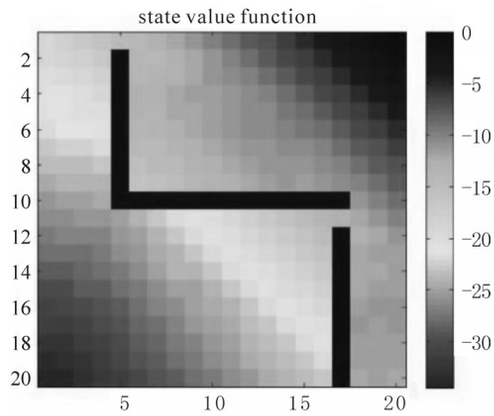

图7 *Q*值表图

Fig.7 *Q*-table

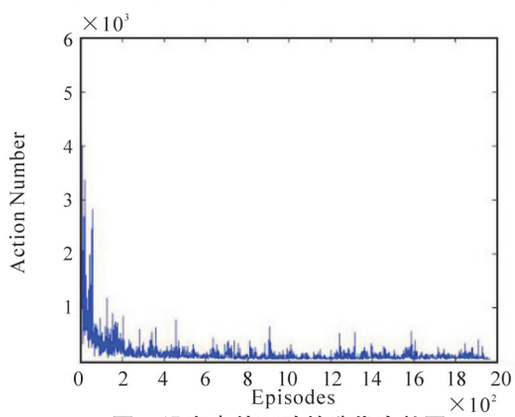

图8 没有事件驱动的迭代次数图

Fig.8 Normal iteration number map without event driven

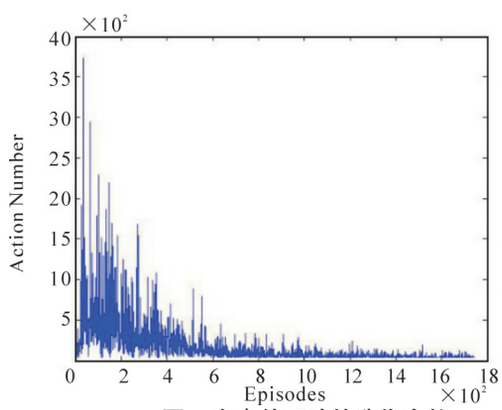

图9 有事件驱动的迭代次数

Fig.9 Iteration number map with event driven

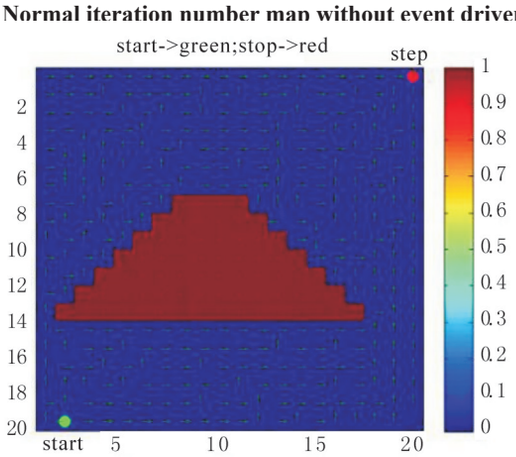

图10 梯形迷宫环境图

Fig.10 Trapezoidal labyrinth environment

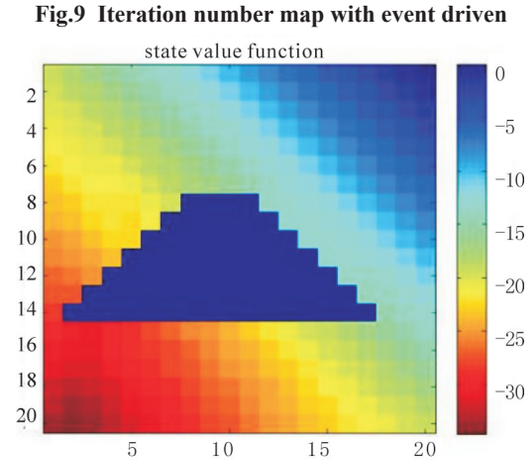

图11 梯形迷宫环境 *Q*值表图

Fig.11 *Q*-table for Trapezoidal labyrinth environment

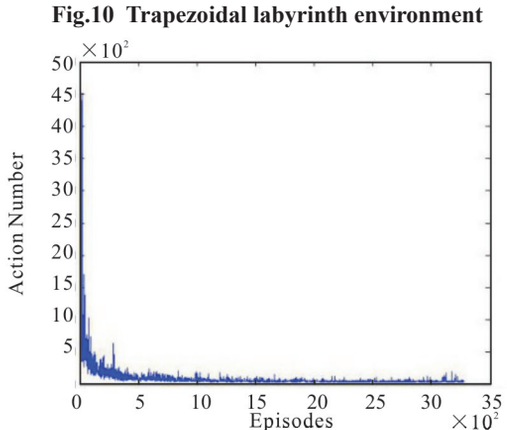

图12 没有事件驱动的迭代次数图

Fig.12 Normal iteration number map without event driven

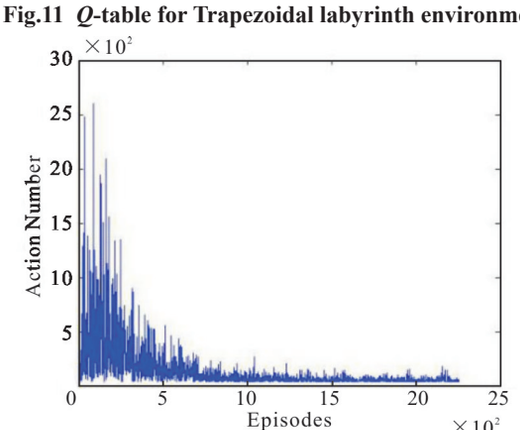

图13 有事件驱动的迭代次数

Fig.13 Iteration number map with event driven

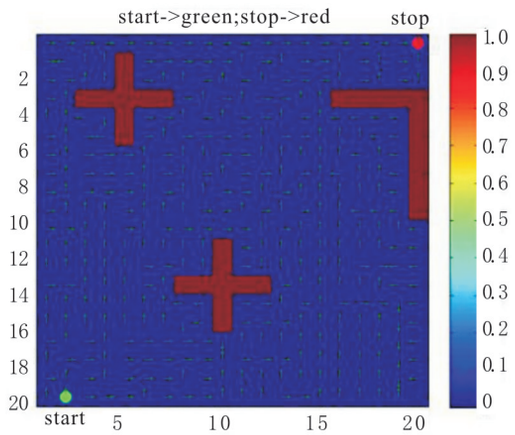

图 14　十字迷宫环境

Fig.14　Cross labyrinth environment

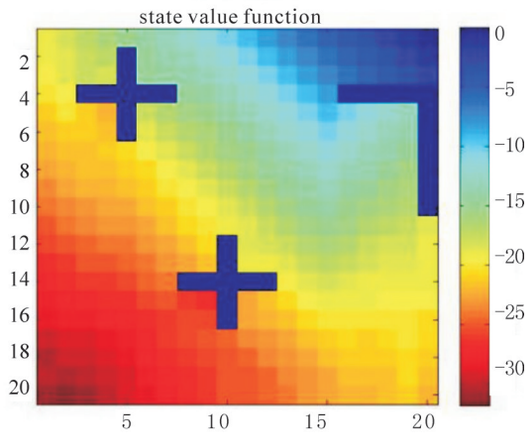

图 15　十字迷宫环境 Q 值表图

Fig.15　Q-table for Cross labyrinth environment

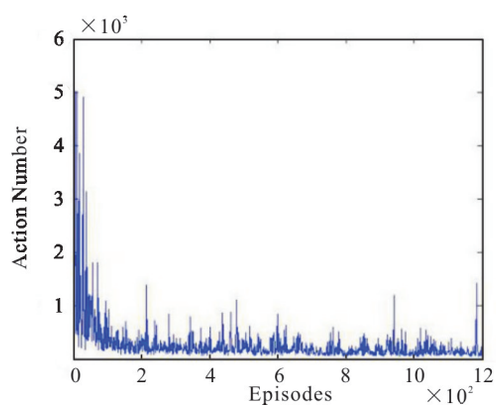

图 16　没有事件驱动的迭代次数图

Fig.16　Normal iteration number map without event driven

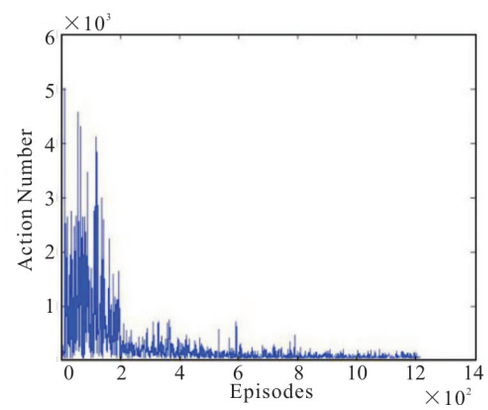

图 17　有事件驱动的迭代次数

Fig.17　Iteration number map with event driven

分别如图 7—图 9、图 11—图 13 和图 15—图 17 所示.从对比实验的迭代次数图中可以发现：引入事件驱动的控制机制后，无人机对于避障动作的策略不需要按照固定的周期来决策；通过事件驱动条件更新无人机的避障行动，有效降低了无人机避障动作决策的频率以及对计算资源的消耗.因此，可以得出本算法具有搜索策略速度快、决策量少的优势.

4　结论

本文提出了一种基于事件驱动的无人机强化学习避障算法，侧重于解决普通强化学习运算次数过多的问题.通过加入事件驱动，使得算法在相同时间内可以明显降低数据的通信次数，并且分析了该算法的主要参数对计算量优化的影响.通过仿真实验说明了该算法可以在学习过程中减少策略遍历次数，解决了强化学习无人机避障算法运算次数过多的问题.

参考文献

[1]　徐亚妮，罗文广，张亮.基于EPGA的四轴飞行器控制系统设计[J].广西科技大学学报，2018，29(3)：50-56.

[2]　陈艳，李春贵，胡波.一种改进的田间导航特征点提取算法[J].广西科技大学学报，2018，29(3)：71-76.

[3]　NARAYANAN V，JAGANNATHAN S. Event-triggered distributed control of nonlinear interconnected systems using online reinforcement learning with exploration [J]. IEEE Transactions on Cybernetics ，2018，48(9)：2510-2519.

[4]　SUTTON R S，BARTO A G. Reinforcement learning：an introduction[M]. Cambridge，MA，USA： MIT Press，1998.

[5]　SAHOO A，XU H，JAGANNATHAN S. Neural network-based adaptive event-triggered control of nonlinear continuous-time systems[C].2013 IEEE International Symposium on Intelligent Control （ISIC），2013：35-40.

[6]　WEN J Y，WANG C，XIE G M. Asynchronous distributed event-triggered circle formation of multi-agent systems[J].Neuro-

computing，2018，295：118-126.
[7] ASTROM K J，BO B. Comparison of periodic and event based sampling for first order stochastic systems[C]. Proceedings of IFAC World Congress，1999，83：301-306.
[8] TABUADA P. Event-triggered real-time scheduling of stabilizing control tasks[J]. IEEE Transactions on Automatic Control，2007，52（9）：1680-1685.
[9] DIMAROGONAS D V，FRAZZOLI E，JOHANSSON K H. Distributed event-triggered control for multi-agent systems[J]. IEEE Transactions on Automatic Control，2012，57（5）：1291-1297.
[10] WANG X F，LEMMON M. Event-triggering in distributed networked control systems[J]. IEEE Transactions on Automatic Control，2011，56（3）：586-601.
[11] SEYBOTH G S，DIMAROGONAS D V，JOHANSSON K H. Event-based broadcasting for multi-agent average consensus [J]. Automatica，2013，49（1）：245-252.
[12] WANG J，ZHANG X M，LIN Y F，et al. Event-triggered dissipative control for networked stochastic systems under non-uniform sampling [J]. Information Sciences，2018，447：216-228.
[13] GUO G，WENS X. Protocol Sequence and control co-design for a collection of networked control systems[J].International Journal of Robust and Nonlinear Control，2015，26（3）：489-508.
[14] GUO G，LU Z B，SHI P. Event-driven actuators：to zero or to hold?[J].International Journal of Robust and Nonlinear Control，2014，24（17）：2761-2773.
[15] GUO G，DING L，HAN Q L. A distributed event-triggered transmission strategy for sampled-data consensus of multi-agent systems[J]. Automatica，2014，50（5）：1489-1496.
[16] DING D R，WANG Z D，DWC H，et al.Observer-based event-triggering consensus control for multiagent systems with lossy sensors and cyber-attacks[J]. IEEE Transactions on Cybernetics，2017，47（8）：1936-1947.
[17] DONKERS M C F，HEEMELS W P M H. Output-based event-triggered control with guaranteed-gain and improved and decentralized event-triggering[J].IEEE Transactions on Automatic Control，2012，57（6）：1362-1376.
[18] FITER C，HETEL L，PERRUQUETTI W，et al. A robust stability framework for LTI systems with time-varying sampling [J] Automatica，2015，54：56-64.
[19] HU L，WANG Z D，HAN Q L，et al. Event-based input and state estimation for linear discrete time-varying systems [J]. International Journal of Conrtrol，2018，91（1）：101-113.
[20] PENG C，HAN Q L. On designing a novel self-triggered sampling scheme for networked control systems with data losses and communication delays[J]. IEEE Transactions on Industrial Electronics，2015，63（2）：1239-1248.
[21] WANG X F，LEMMON M D. Self-triggered feedback control systems with finite-gain stability[J].IEEE Transactions on Automatic Control，2009，54（3）：452-467.
[22] ZOU L，WANG Z D，ZHOU D H. Event-based control and filtering of networked systems：a survey[J]. International Journal of Automation & Computing，2017，14（3）：239-253.
[23] VAMVOUDAKIS K G，FERRAZ H. Model-free event-triggered control algorithm for continuous-time linear systems with optimal performance [J]. Automatica，2018，87：412-420.

Research on obstacle avoidance for UAV using reinforcement learning based on event driven

TANG Bowen[a]， WANG Zhiwen[*b]， HU Zhenhuan[a]
（a. School of Electrical and Information Engineering， Guangxi University of Science and Technology， Liuzhou 545006，China; b. School of Computer Science and Telecommunication Engineering，Guangxi University of Science and Technology， Liuzhou 545006， China）

Abstract：The reinforcement learning method is widely used in the research of obstacle avoidance，and it needs to consume a large amount of computing resources. This paper proposes an event-driven drone reinforcement learning obstacle avoidance algorithm. By adding event-driven triggering mecha

（下转第117页）

[8] 熊维玲，甘桦源.（3+1）维Jimbo-Miwa方程的非行波解[J].广西科技大学学报，2017，28（1）：12-18.
[9] 施业琼.（2+1）维Davey-Stewatson II方程的精确解[J].广西科技大学学报，2015，26（1）：96-102.
[10] CALOGERO F，DEGASPERIS A. Nonlinear evolution equations solvable by the inverse spectral transform[J]. II Nuovo Cimento della Società italiana di Fisica-B： General Physics，1977，39（1）： 1-54.
[11] ZHANG J F，MENG J P. New localized coherent structures to the（2+1）-dimensional breaking soliton equation[J]. Physics Letters A，2004，321（1）：173-178.
[12] ZHANG S.A generalized new auxiliary equation method and its application to the（2+1）-dimensional breaking soliton equations[J].Applied Mathematics and Computation，2007，190（1）：510-516.
[13] HASIBUN N，FARAH A A. New generalized and improved (G'/G)-expansion method for nonlinear evolution equations in mathematical physics[J]. Journal of the Egyptian Mathematical Society，2014，22（3）：390-395.

Using extended expansion method to obtain exact solutions of （2+1）-dimensional breaking soliton equation

LIAO Ganjie， HUANG Liwei*， CHEN Xian， GUO Yanfeng
（College of Science， Guangxi University of Science and Technology， Liuzhou 545006， China）

Abstract：Using the extended -expansion method and new auxiliary equations，the new exact solutions of（2+1）-dimensional breaking equation are obtained by the homogeneous balance method. And some exact solutions of（2+1）-dimensional breaking soliton equation are given. In addition，the corresponding numerical simulation images of some solutions are given and analyzed.

Key words： breaking soliton equation; homogeneous balance; exact solution; （G'/G）-expansion method

（责任编辑：张玉凤）

（上接第102页）

nism to reinforcement learning，the optimal path can be found while the UAV action decision is reduced，which can ensure the performance and reduce the communication frequency of the system. The simulation results show that the algorithm can reduce the consumption of computing resources in the learning process and achieve obstacle avoidance tasks while significantly accelerating the convergence rate.

Key words：event driven; reinforcement learning; obstacle avoidance; Q-learning

（责任编辑：黎　娅）