中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

一种新的基于强化学习改进SAR的无人机路径规划

周文娟, 张超群, 汤卫东, 易云恒, 刘文武, 秦唯栋

引用本文:
周文娟,张超群,汤卫东,易云恒,刘文武,秦唯栋. 一种新的基于强化学习改进SAR的无人机路径规划[J]. 控制与决策, 2024, 39(4): 1203−1211.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.1209

一种新的基于强化学习改进SAR的无人机路径规划

周文娟[1], 张超群[1,2†], 汤卫东[1], 易云恒[1], 刘文武[1], 秦唯栋[1]

(1. 广西民族大学 人工智能学院，南宁 530006；
2. 广西混杂计算与集成电路设计分析重点实验室，南宁 530006)

摘 要: 搜索和救援优化算法(SAR)是2020年提出的模拟搜救行为的一种元启发式优化算法,用来解决工程中的约束优化问题. 但是,SAR存在收敛慢、个体不能自适应选择操作等问题,鉴于此,提出一种新的基于强化学习改进的SAR算法(即RLSAR). 该算法重新设计SAR的局部搜索和全局搜索操作,并增加路径调整操作,采用异步优势演员评论家算法(A3C)训练强化学习模型使得SAR个体获得自适应选择算子的能力. 所有智能体在威胁区数量、位置和大小均随机生成的动态环境中训练,进而从每个动作的贡献、不同威胁区下规划出的路径长度和每个个体的执行操作序列3个方面对训练好的模型进行探索性实验. 实验结果表明,RLSAR比标准SAR、差分进化算法、松鼠搜索算法具有更高的收敛速度,能够在随机生成的三维动态环境中成功地为无人机规划出更加经济且安全有效的可行路径,表明所提出算法可作为一种有效的无人机路径规划方法.

关键词: 强化学习；搜索与救援优化算法；异步优势演员-评论家算法；路径规划；路径调整；无人机

中图分类号: TP301.6 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.1209

引用格式: 周文娟,张超群,汤卫东,等. 一种新的基于强化学习改进SAR的无人机路径规划[J]. 控制与决策, 2024, 39(4): 1203-1211.

A novel modified search and rescue optimization algorithm based on reinforcement learning for UAV path planning

ZHOU Wen-juan[1], *ZHANG Chao-qun*[1,2†], *TANG Wei-dong*[1], *YI Yun-heng*[1], *LIU Wen-wu*[1], *QIN Wei-dong*[1]

(1. College of Artificial Intelligence，Guangxi Minzu University，Nanning 530006，China；2. Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis，Nanning 530006，China)

Abstract: The search and rescue optimization algorithm (SAR) proposed in 2020 is a meta-heuristic optimization algorithm. It simulates the search and rescue behavior, which is used to solve constrained engineering optimization problems. However, the SAR has slow convergence and its individuals can not adaptively select operations. A modifed version of the SAR based on reinforcement learning, namely RLSAR, is proposed, which redesigns the local search and global search of the SAR, and adds path adjustment operation. The asynchronous advanced actor critic algorithm (A3C) is used to train the reinforcement learning model so that the SAR individuals acquire the ability to adaptively select operators. All agents are trained in a dynamic environment in which the number, location and size of threat areas are randomly generated, and then exploratory experiments are conducted on the trained model from three aspects: The contribution of each action, the path length planned under different threat areas, and the execution sequence of each individual. The results show that the RLSAR has higher convergence speed than the standard SAR, the differential evolution algorithm and the squirrel search algorithm. Furthermore, it can successfully plan a more economical, safe and effective feasible path for an unmanned aerial vehicle (UAV) in a randomly generated three-dimensional dynamic environment, which shows that the proposed algorithm can serve as an effective path planning method for UAVs.

Keywords: reinforcement learning；search and rescue optimization algorithm；asynchronous advantage actor-critic algorithm；path planning；path adjustment；unmanned aerial vehicle

0 引 言

近年来无人机(unmanned aerial vehicle, UAV)广泛应用于覆盖规划[1]、多目标跟踪[2]、测绘[3]和监视[4]等领域. 在实际飞行过程中,无人机需要在不穿过威胁区的前提下通过有效的路径规划方法获取可行、高效的飞行方案. 有学者提出A*[5]、Dijsktra[6]和RRT[7]

收稿日期: 2022-07-07；**录用日期:** 2023-01-02.
基金项目: 国家自然科学基金项目(62062011)；广西民族大学研究生创新计划项目(gxun-chxs2021057).
责任编委: 侯忠生.
†通讯作者. E-mail: 25713893@qq.com.

等优化算法解决无人机路径规划问题,但是,这类算法只能给出可行解. 随着优化领域的发展,群智能优化算法如遗传算法[8]、粒子群优化算法[9]、果蝇优化算法[10]和灰狼优化算法[11]等相继被提出,这些智能算法广泛应用于路径规划问题[12-15]. 群智能优化算法通过模拟个体间互相交流的方式进行寻优,具备一定的智能程度,但是,种群内的个体只能以特定的算子选择结果进行更新. 这种策略没有充分利用个体状态信息,导致迭代过程中算法无法自适应选择算子,从而限制了算法性能.

搜索和救援优化算法(search and rescue optimization algorithm, SAR)是Shabani等[16]于2020年提出的一种元启发式优化算法,用来解决工程中的约束优化问题. SAR受人类搜救活动的搜索策略的启发,通过对搜救人员行为的建模而形成. 该算法包含社会阶段和个人阶段两个算子,前者用来进行局部搜索,后者用来进行全局搜索,两个阶段交替进行. 与上述元启发式算法相比,SAR个体利用与其他个体间的位置关系更新自身,这种新的优化思想具有灵活、简单、易于实现且不易陷入局部最优等特点,在UAV的路径规划中很少被使用. 然而,它也存在算子选择的非适应性问题,即无法根据种群个体在迭代中所处状态自适应地调整合适的算子指导自身更新.

强化学习(reinforcement learning, RL)是机器学习的一个分支[17]. 不同于监督学习算法,强化学习的本质是智能体(agent)的“试错”和更新,其通过环境的评价性反馈指导行动,从而获得最多的奖赏[18],这更接近于人类的学习过程. 自强化学习被提出以来,其在路径规划和导航问题上得到越来越多的应用. 如Ai等[19]提出了一种基于强化学习的海上搜索和救援自主覆盖路径规划模型,并设计了具有多个约束的奖励函数来指导船舶的导航动作,最终得出比其他路径规划算法覆盖率更高,且长度更短的路径;Gismondi等[20]研究了类独轮车移动机器人的路径规划问题,并通过强化学习来控制独轮车避免碰撞;Qu等[21]提出了一种基于强化学习的灰狼优化算法来解决无人机在复杂环境下的路径规划问题.

为了解决SAR个体不能根据种群状态自适应地选择算子的问题,提出一种基于强化学习的搜索和救援算法(reinforcement learning-based search and rescue optimization algorithm, RLSAR),将SAR中的个体作为强化学习的智能体,代价函数值作为评价信号,采用异步优势演员评论家算法(asynchronous advantaged actor-critic algorithm, A3C)[22]训练强化学习网络指导SAR的个体选择操作,针对每个智能体设置局部搜索、全局搜索和路径调整3种操作,新建一种具备泛化能力的强化学习模型,使得SAR在路径规划中可自适应地为个体选择操作,从而提升SAR的整体性能.

1 标准SAR

SAR在搜索过程中,团队成员以收集、保存和更新线索的形式寻找最优解. SAR同时保存搜救人员(个体)当前的位置X和其留下的线索位置M,如下式所示,它是SAR优化算法的重要部分. 矩阵C、M和X在每个个体搜索时均会更新,即

$$C=\begin{bmatrix}X\\M\end{bmatrix}=\begin{bmatrix}X_{11}&\cdots&X_{1D}\\\vdots&\ddots&\vdots\\X_{N1}&\cdots&X_{ND}\\M_{11}&\cdots&M_{1D}\\\vdots&\ddots&\vdots\\M_{N1}&\cdots&M_{ND}\end{bmatrix}. \tag{1}$$

其中:N为搜救小组成员的个数,D为问题的维数,X_{N1}为第N个搜救人员第1维的位置信息,M_{1D}为第1个搜救人员留下的线索中第D维的位置信息.

SAR优化算法包含社会阶段(social phase)和个人阶段(individual phase)两个主要的算子. 前者通过社会效应(social effect)的参数SE来控制局部搜索与全局搜索的平衡,后者负责局部搜索,社会和个人阶段搜索方案均基于线索矩阵C来更新.

1.1 社会阶段

每个搜救人员随机选取一个线索,若被选取的线索比当前位置的线索更重要,则围绕被选取的线索搜索新的解;否则,搜索将围绕当前位置进行. 个体在社会阶段按照下式更新:

$$X'_{ij}=\begin{cases}U+r_1\times(X_{ij}-C_{kj}),\ r_2<\text{SE or } j=j_r;\\X_{ij},\ \text{otherwise.}\end{cases}$$

$$U=\begin{cases}C_{kj},\ f(C_k)\leqslant f(X_i);\\X_{ij},\ \text{otherwise.}\end{cases}$$

$$i,k=1,2,\ldots,N,\ j=1,2,\ldots,D. \tag{2}$$

其中:X'_{ij}为更新后第i个搜救人员在搜索空间第j维上的值;U为搜救起点;f为目标函数;C_{kj}为第k个线索第j维值;文献[17]建议SE取0.05,且r_2是$[0,1]$上均匀分布的随机数;j_r为介于$[1,D]$间的随机整数,这保证了X'_i与X_i至少在一个维度上不同;r_1为介于$[-1,1]$间均匀分布的随机数,对于不同的j,r_1只随机一次.

1.2　个人阶段

搜救人员以当前位置为起点进行搜索,通过2个随机线索的差向量确定下一次的搜索方向和行进距离. 第i个搜救人员的新位置按照下式更新:

$$X_i' = X_i + r_3 \times (C_k - C_m). \tag{3}$$

其中: k和m为介于$[1, 2N]$间的随机整数且$i \neq k \neq m$, r_3为在区间$[0, 1]$上服从均匀分布的随机数.

1.3　越界处理

若部分SAR个体在更新时出现越界的情况,即其新位置在搜索空间外,则需要对这些个体进行越界处理. SAR按照下式对越界的个体进行处理:

$$X_{i,j}' = \begin{cases} \dfrac{X_{i,j} + \mathrm{ub}_j}{2}, & X_{i,j}' > \mathrm{ub}_j; \\ \dfrac{X_{i,j} + \mathrm{lb}_j}{2}, & X_{i,j}' < \mathrm{lb}_j. \end{cases}$$
$$i = 1, 2, \ldots, N,\ j = 1, 2, \ldots, D. \tag{4}$$

其中ub_j和lb_j分别为搜索空间中第j维中的最大值和最小值.

通常情况下最适合个体的更新策略会随着个体的位置、当前最优值和当前所处的迭代阶段变化. 然而,在标准SAR的更新策略中,社会阶段和个人阶段交替进行,这使得该算法不能以最高效的方式收敛. 为了令SAR的算子具备自主选择行动的能力,将SAR与强化学习思想相结合,利用智能体对环境的反馈来调整SAR个体动作策略.

2　强化学习与A3C

强化学习主要由智能体、环境、状态、动作和奖励组成. 由于智能体与环境的交互方式与人类类似,可认为强化学习是一套通用的学习框架,可用来解决通用人工智能的问题. 本文采用A3C强化学习模型对SAR个体进行训练,对路径规划问题构建了第3.1节所描述的数学模型.

A3C从改进网络结构、Critic评估方式和异步训练方式等方面提升传统Actor-Critic结构的性能. 其创建多个并行的环境,令多个拥有副结构的智能体同时在这些并行环境上更新主结构的参数. 并行的智能体们互不干扰,而主结构的参数更新受到副结构提交更新的非连续性干扰, 以降低过拟合的概率,提高训练的收敛性能.

3　RLSAR

3.1　数学模型

3.1.1　环境威胁模型

在连续空间下建立如图1所示的环境威胁模型,该模型满足以下条件.

1) S为无人机路径规划的起点, E为终点.

2) 将环境中对无人机安全产生影响的物体抽象为球状威胁区,当无人机的某一段航迹经过威胁区时,无人机安全受到威胁,如雷达的监测和山脉的碰撞等.

3) T_k为第k个威胁区的中心, r_k为半径, $k = 1, 2, \ldots, K$, K为最大威胁区个数.

4) 地图在X、Y、Z轴上的长度分别为L、W、H. 地图区域中包含所有威胁区的中心. S点坐标为$(0, W/2, H/2)$, E点坐标为$(L, W/2, H/2)$.

5) 路径点P_j为SAR个体在地图中所表示路径上的第j个节点. 整个飞行路径由起点、路径点和终点顺次连接而成,记为$\{S, P_1, P_2, \ldots, P_D, E\}$,其中$D$为SAR中个体的个数,在三维环境中, $D \times 3$为SAR个体的维度.

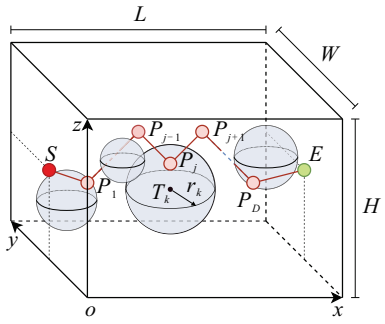


图1　环境威胁模型

3.1.2　代价函数

假设无人机以恒定的速度飞行,且不考虑空气阻力. 在UAV飞行过程中,考虑无人机的燃料成本和威胁成本,且UAV的燃料成本C_{fuel}与UAV的航迹长度成正比,可按照下式计算:

$$C_{\mathrm{fuel}} = \sum_{j=1}^{D} l_j, \tag{5}$$

其中l_j为连接P_{j-1}与P_j的线段长度.

当规划出的路径线段经过威胁区时,引入威胁代价函数C_{threat},有

$$C_{\mathrm{threat}} = \sum_{j=1}^{D} \sum_{k=1}^{K} l_{j,k}', \tag{6}$$

其中$l_{j,k}'$为线段$\overline{P_{j-1}P_j}$在第k个威胁区中的长度.

在无人机路径规划中,问题的优化目标为最小化C_{fuel}与C_{threat}之和,实现总成本C_{total}最低,即

$$\min\ C_{\mathrm{total}} = C_{\mathrm{fuel}} + C_{\mathrm{threat}}. \tag{7}$$

3.2　RLSAR的结构

所提出RLSAR整体结构如图2所示. 在RLSAR结构中, A2C网络输入层I的维度为201,隐藏层H_1、H_2的维度分别为32和16. A2C网络输出有两个: 1) Actor网络输出的动作集合A,维度为Action的个数

(为3); 2) Critic网络对于Actor网络的评价值C,维度为1.

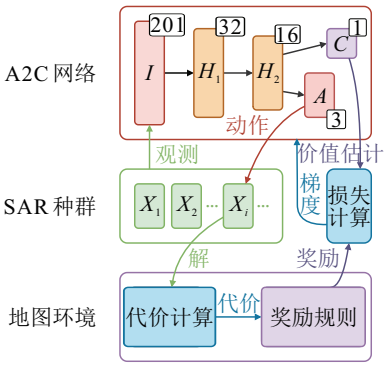


图 2 RLSAR的结构

RLSAR种群个体的行动策略得到一个A2C网络的指导,每个个体将对种群中其他所有个体的观测输入至A2C网络中,然后通过Actor网络给出的动作来判断个体执行社会阶段、个人阶段或路径调整. 执行后,个体成为一个新的可行解,地图环境对这个可行解进行飞行成本C_{total}的计算以判断奖励值. 同时,Critic网络输出对Actor的评价,这一评价值与实际奖励进行对比以计算网络损失,然后再反向传播梯度信息,调整网络参数.

下文将详细描述RLSAR中的状态、动作的设置以及奖励函数设计.

3.3 状态设置

为了保证强化学习训练出的网络具备泛化能力,地图环境对于SAR个体是未知的,且威胁区的数量和位置是随机生成的,智能体(SAR种群的单个个体)通过观测SAR个体的状态来判断和进行后续行为. 智能体对个体状态的观测分为3个部分: 1) SAR所有个体与当前个体的相对位置; 2) SAR所有个体的目标函数值与当前个体的差向量; 3) SAR当前已经历的迭代次数与最大迭代次数的比值.

3.4 动作设置

3.4.1 局部搜索与全局搜索

SAR的社会阶段负责局部搜索(local exploration),通过获取同伴个体的信息来更新当前个体;SAR的个人阶段负责全局搜索(global exploration),通过选取任意两个个体作差形成一个移动向量,引导被更新的个体进行移动. RLSAR保留两者,分别作为RLSAR的局部搜索算子和全局搜索算子.

3.4.2 路径调整

无人机在飞行时会有多次转弯,导致路径出现弯曲,造成不必要的燃料损失. 如文献[23]中的修正变异算子,其提出了一种实时路径调整(adjustment)策略,采用下式对个体所表示路径进行随机拉直:

$$P_j' = \frac{P_{j-1} + P_{j+1}}{2}. \tag{8}$$

从点P_1与点P_D的路径上随机选取两个点P_{j-1}和P_{j+1},创造一个中间路径点P_j',若P_j'未落入威胁区,则由式(8)将原来路径中的P_j变更为P_j',这种方法使得从P_{j-1}点到P_{j+1}点的路径更短,避免飞行路线中不必要的弯曲.

3.5 奖励设计

在强化学习任务中,合理的奖励对于强化学习网络有至关重要的影响. 在群智能优化算法中,每个个体在做出符合自身提升行动的同时,还要考虑全局最优的寻找. 前者在更加合理的范围进行全局搜索,使得群体整体得到优化,后者重视全局最优附近的局部搜索,两者均为SAR重要的组成部分. 奖励的设置分别从个体提升、全局最优单步提升、全局最优与目标值差距3个方面进行个体行动的价值判断.

在训练过程中,智能体移动后的位置若优于移动前,则智能体会得到奖励;否则,会受到惩罚(奖励为负值). 这种奖励设置方法可有效避免智能体为得到最大奖励而进行无限次的循环,从而无法得到更好的训练结果.

3.5.1 个体提升

将第i个个体移动前后的目标函数值分别记为$f(X_i)$和$f(X_i')$. 若$f(X_i) < f(X_i')$,则表明智能体移动后没有变好,智能体将会得到一个-1的奖励;否则,按照下式计算奖励:

$$R_1 = [f(X_i) - f(X_i')] \times \frac{f(X_i') - f_w}{f_b - f_w} \times \frac{1}{N}. \tag{9}$$

其中: R_1为个体提升的奖励, f_w为当前最差个体的目标函数值, f_b为当前全局最优个体的目标函数值, N为智能体的数量.

此外,设置一个惩罚项以避免智能体过早调整路径,设置该惩罚项的原因是路径调整算子引导个体在自身周围的空间中搜索更优解. 在种群未充分收敛至最优解附近时,若过早进行路径调整则会增加种群陷入局部最优的概率. 若个体前期进行路径调整,则按照下式对其进行惩罚:

$$R_2 = \begin{cases} -3 \times \dfrac{80 - t}{80}, & \text{若路径调整且 } t \leqslant 80; \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

其中: R_2为智能体进行路径调整操作受到的惩罚, t为智能体当前迭代的次数.

3.5.2 全局最优单步提升

若个体在更新后改变了全局最优,新的全局最优记为f_b',则全局最优的单步提升奖励R_3按照下式计

算:

$$R_3 = (f_b - f_b') \times \frac{1}{N}. \tag{11}$$

3.5.3 全局最优与目标值差距

在全局最优确定后,通过下式计算全局最优与目标值的差距奖励R_4:

$$R_4 = (F - f_b) \times \frac{1}{N} \times \frac{1}{T_{\max}}. \tag{12}$$

其中:$T_{\max}$为最大迭代次数,F为一个路径规划问题的目标函数值.

对整体奖励R的计算定义如下:

$$R = (R_1 + R_2) \times 0.5 + R_3 + R_4. \tag{13}$$

3.6 计算复杂度分析

根据设计好的动作和奖励,分析SAR和RLSAR的复杂度,由文献[16]可知,SAR的计算复杂度主要由迭代次数T_1、个体数量N以及维度D构成,即$O(\text{SAR}) = O(T_1 \times N \times D)$.

RLSAR的复杂度分析分为训练和调用两部分.在训练过程中,每回合的最大迭代次数为T_2,共训练E回合,改进后的SAR在每个回合均进行一次完整的寻优过程,由此得到训练复杂度为$O(\text{RLSAR}) = O(E \times T_2 \times N \times D)$.

在针对实际问题时,通常使用训练好的模型进行优化.在模型调用过程中,A3C网络计算复杂度为$O(1)$,路径调整操作计算复杂度为$O(1)$,由RLSAR结构(如图3所示)可知,RLSAR的计算复杂度为SAR更新、A3C网络更新与路径调整操作的计算复杂度之和:$O(\text{RLSAR}) = O(\text{SAR}) + O(1) + O(1) = O(\text{SAR}) = O(T_1 \times N \times D)$,与SAR的计算复杂度同阶.

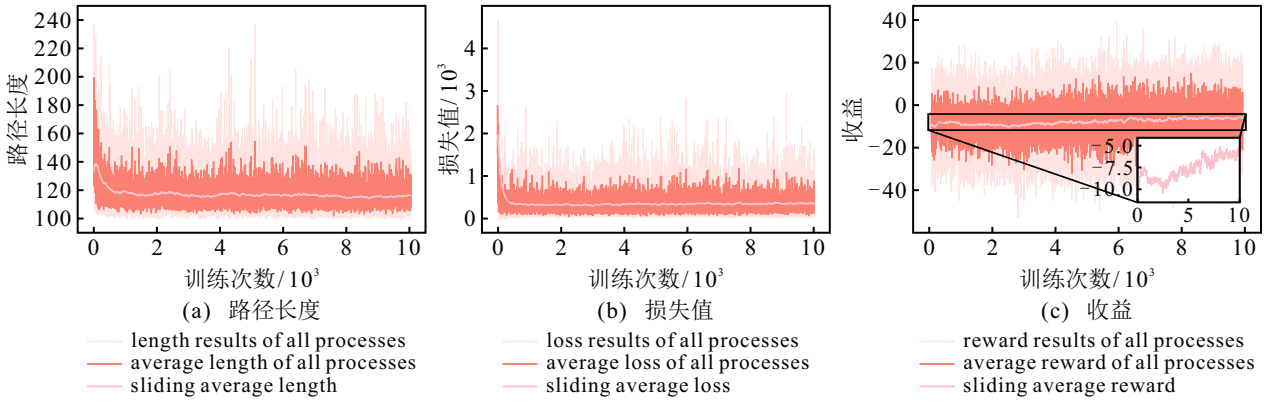


图3　RLSAR训练结果

4 实验与结果分析

4.1 超参数设置和性能指标

Shabani等[16]在提出标准SAR后,分别用13个基准约束函数和包含10维、30维的18个可扩展约束函数做测试,前者设定个体数量(种群规模)为20,后者个体数量分别设为25、40,令SAR独立运行,计算标准差.由计算结果可知,在高维环境中适当增加个体数量后,得到与较少个体相似的结果,因为种群规模会影响求解的精度,但是,种群规模达到一定数量后其结果已趋向稳定,再增加个体,其求解效果影响不大或作用不明显,反而会增加计算代价.

为了测试RLSAR的性能,分别使用RLSAR、SAR、差分进化算法(differential evolution, DE)[24]和松鼠搜索算法(squirrel search algorithm, SSA)[25]求解路径规划问题.各算法和地图的参数如表1所示.其中:起点与终点间的最短距离为100,无人机路径规划目标理想的最优值为100,所有威胁区均匀分布于搜索空间内,半径服从$r \sim N(25, 13)$的正态分布,威胁区数量在区间$[4, 8]$中随机生成.

表1　算法和地图的参数设置

名称	参数	参数说明	设置值
SAR/RLSAR	pop	种群大小	8
	T_1	测试最大迭代次数	300
DE	pop	种群大小	8
	diff strategy	差分策略	rand/1
SSA	pop	种群大小	20
地图	L	长度	100
	W	宽度	100
	H	高度	100
	D_M	维度	3
	D	解的维度	24
	K	威胁区数量	4 ~ 8
	N	节点数量	8
训练	learning rate	学习率	10^{-4}
	action num	动作数量	3
	E	训练最大回合	10 000
	F	目标最优值	100
	l	前视步数	100
	T_2	训练最大迭代次数	100
	processes	线程数量	4

4.2 实验结果分析

4.2.1 RLSAR模型训练结果

将RLSAR模型置于16 GB RAM、Intel Core i7 10 700 2.90 GHz CPU、Windows 11系统的个人计算机上训练,集成开发环境是PyCharm, Python 3.7. 模型训练结果如图3所示.

图3中,由浅至深3个颜色分别为4个线程训练结果的叠加曲线、平均值和平均值的滑动平均曲线. 图3(a)表明RLSAR规划的路径长度最终稳定在115左右;图3(b)显示损失函数值(loss)最终趋于稳定;图3(c)显示回报函数值的变化过程,在该图的放大图中可以清晰地看出,回报值在波动中上升并趋于稳定,表明RLSAR模型的训练过程是有效的,RLSAR具备了自主选择执行阶段的能力.

4.2.2 路径规划结果

表2为威胁坐标生成3个由简至繁的实例,观察4种算法规划出的路径结果.

表 2 3个实例中各威胁区的坐标设置

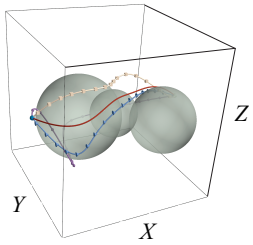
实例	坐标	半径
1	(32, 55, 43)	27
	(50, 43, 48)	15
	(77, 28, 46)	21
2	(42, 46, 43)	18
	(20, 26, 47)	20
	(80, 69, 46)	12
	(77, 33, 45)	19
	(21, 73, 43)	16
3	(41, 52, 39)	20
	(33, 26, 27)	18
	(83, 64, 50)	16
	(63, 76, 41)	16
	(77, 33, 45)	14
	(24, 77, 44)	20
	(43, 52, 43)	17

所有路径均采用3次B样条[26]进行平滑. 由于飞行空间3个维度的取值连续,规划出的路径中可能存在距离过近的节点,3次B样条插值会使得平滑路径与原折线路径差别过大. 故当节点$P_i(x_i, y_i, z_i)$与$P_{i+1}(x_{i+1}, y_{i+1}, z_{i+1})$间的距离$l_{i+1}$小于固定值lp时,采用下式处理节点:

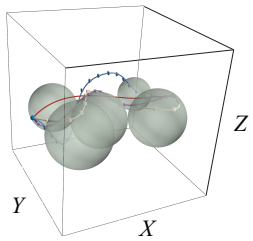
$$w_i' = \frac{w_i + w_{i+1}}{2},\ w = x, y, z. \tag{14}$$

其中w_i'为处理后P_i点在各维度上的坐标值. $P_i'(x_i', y_i', z_i')$为第i个路径节点. 根据问题复杂度,取lp = 15. 预处理后,删除路径中原有的2个节点P_i和P_{i+1},将新的节点P_i'加入路径中.

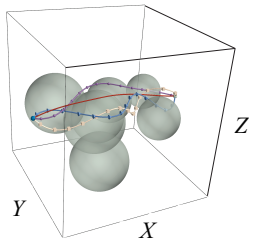
图4(a)为4种算法在实例1中的路径规划结果. 由图4(a)可见:RLSAR规划出的路径长度明显短于其余3种算法,最接近最优解. DE和SAR找到了较优的路径,虽然SSA使用了20个个体进行规划,但是,表现依然不如有8个个体的RLSAR. 图4(b)为实例2的规划结果. 由图4(b)可见:RLSAR能够规划出一条平滑且更加经济的路径. 在增加威胁区数量和排列复杂度后,SSA表现有所提升,而SAR和DE的性能有所下降. 图4(c)为实例3的规划结果. 由图4(c)可见:威胁区的复杂度继续增加,RLSAR规划出的路径依然优于其他3种算法,这体现了RLSAR模型的有效性和稳定性. 虽然SSA、DE和SAR的性能没有出现明显的下降,但是,它们规划出来的路径均存在多处非必要的弯曲.



(a) 实例1斜视图



(b) 实例2斜视图



(c) 实例3斜视图

图 4 4种算法对实例1～实例3路径规划结果

4.2.3 收敛速度与规划结果对比

利用训练好的RLSAR模型,在有3、5、7个威胁区的地图上分别使用上述4种算法进行求解,实验结果如图5所示. 由图5可见:在不同威胁区数量下,SAR、DE和SSA均多次陷入局部最优,虽然能够跳出,但是,它们的结果依然更差,收敛速度也更慢. RLSAR规划出的结果最终稳定在110,在整个过程中未陷入局部最优,表明RLSAR具备更加优秀的算子选择策略和收敛性能.

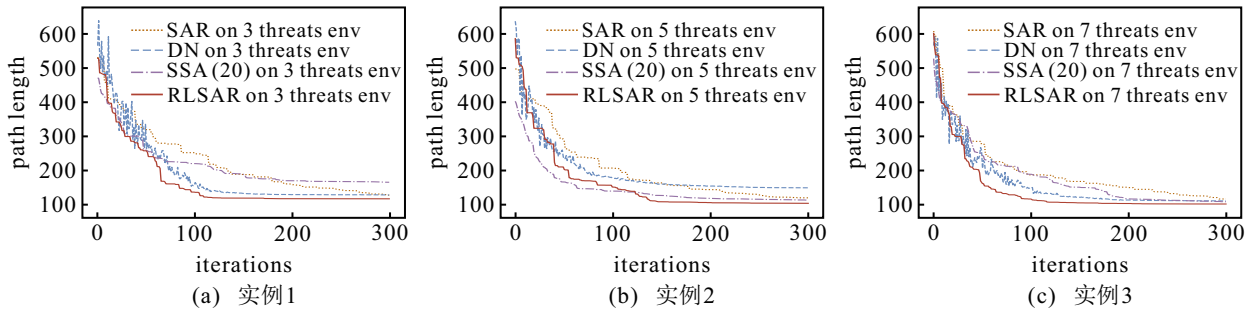


图 5　4种算法在实例1～实例3下的收敛曲线

4种算法在3个实例下规划出的路径长度和所需运行时间数据如表3所示. 由表3可见, RLSAR规划用时比标准SAR高出10%～20%, 这是因为RLSAR在个体的每次更新中需要由强化学习网络计算当前需要选择的最优操作. 但是, 路径长度提升了8%～13%, 在实际的路径规划中可大幅降低无人机飞行成本.

表 3　4种算法规划出的路径长度和运行时间统计

实例	项目	算法名称			
		SAR	DE	SSA	RLSAR
实例1	最优解	130.242	130.943	167.922	**119.596**
	时间/s	1.472	1.816	3.576	1.648
实例2	最优解	119.368	148.715	112.359	**103.312**
	时间/s	1.627	2.673	3.956	1.929
实例3	最优解	117.145	111.550	108.980	**102.354**
	时间/s	2.219	3.475	5.220	2.470

4.3　实验再探究

4.3.1　不同威胁区数量下RLSAR和SAR规划结果

在威胁区位置和数量随机的环境下独立进行100次路径规划, 结果如图6所示. 图6中, RLSAR规划出的路径长度明显优于SAR. SAR的性能随着威胁区数量的升高而降低, RLSAR的性能却有所提升, 这表明SAR对搜索空间复杂度的提升更加敏感, 而RLSAR具有更好的在全局与局部搜索间切换的策略, 能够降低搜索空间复杂度对其造成的负面影响. 整体而言, RLSAR能够为UAV规划一条经济可行的飞行路径且不受威胁区数量的影响.

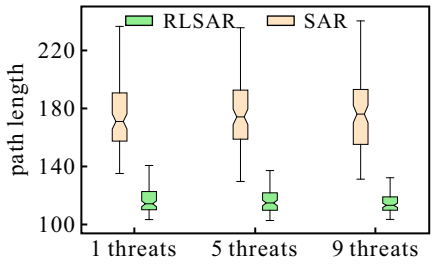


图 6　RLSAR和SAR在不同威胁区数量下的规划结果

4.3.2　消融实验中每个动作对RLSAR的贡献

上述实验验证了RLSAR的稳定性, 为了测试RLSAR动作设计合理性, 分别去掉RLSAR中的不同动作, 对100个随机路径规划问题进行求解, 结果如图7所示. 图7中: 去掉全局搜索后, 算法性能有所下降, 但是, 局部搜索和路径调整操作可使得算法找到全局最优. 去除路径调整动作后, 算法规划出的路径长度波动幅度最大; 去除局部搜索后, 算法能够得出接近原算法的路径结果. 之所以得出上述结果, 是因为路径调整是针对路径规划问题中解空间的特殊性设计的, 本质上是一种特殊的局部搜索, 能够降低SAR中社会阶段对最终结果的影响. 路径调整操作使得解个体进行有向寻优, 对于全局最优而言, 一个解在经历路径调整后趋向更优的可能性较大, 故RLSAR往往会倾向于选择路径调整操作.

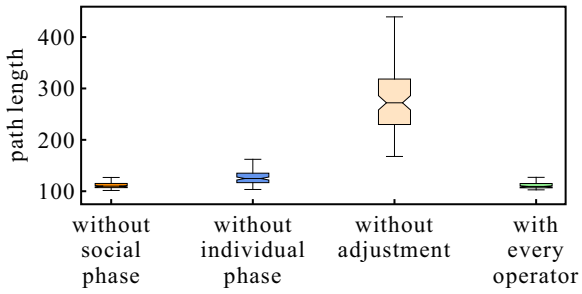


图 7　消融实验结果

为了验证强化学习对于在SAR个体操作选择中产生的积极作用, 训练了不包含路径调整操作的模型RLSAR (wa), 并在最复杂的实例3环境下进行测试, 结果如图8所示.

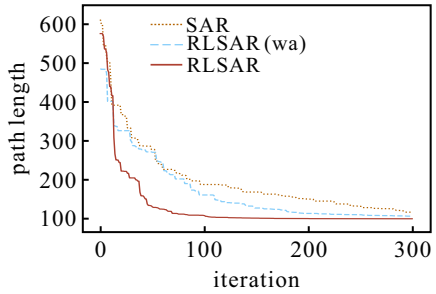


图 8　无路径调整策略的路径规划结果对比

图8中: RLSAR (wa)收敛速度高于标准SAR, 加入路径调整策略的RLSAR收敛速度和最终规划出的路径长度均优于SAR和RLSAR (wa), 表明所采取的路径调整策略对强化学习指导下的SAR性能有明显提升.

4.3.3 RLSAR算法中各个体执行的操作序列分析

图9为8个个体在100次迭代中选择的操作以及全局最佳个体的标识(表示为实心). 3个操作用不同的符号进行区分, 其中, 局部搜索、全局搜索和路径调整分别用△、◯和□表示.

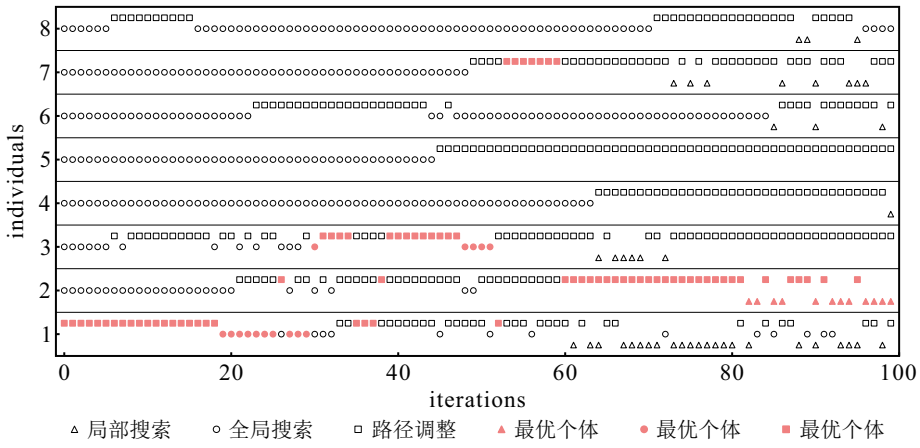


图 9 RLSAR中个体执行的操作序列

由图9可见: RLSAR在搜索前期, 个体更多地探索以快速收敛; 在搜索后期, 个体更多地趋向于局部开采以达到全局最优. 且可以看出, 当个体1选择进行局部搜索而非路径调整时, 个体2选择了进行路径调整, 最优个体在此后也从个体1变为了个体2. 本文通过结合强化学习对SAR进行改进, 成功使得个体能够自适应地选择对自身更有利的动作, 极大地提高了RLSAR的收敛性能.

5 结 论

针对SAR的算子不能自适应根据个体的状态调整个体更新策略从而导致算法收敛慢的问题, 提出了采用强化学习指导SAR更新的思想, 将SAR的每个个体作为强化学习中的智能体, 通过根据SAR个体状态对其进行奖惩, 指导智能体从局部搜索、全局搜索和路径调整中选择对当前个体最有利的动作. 通过设计威胁区数量和位置随机生成的环境地图, 将所提出RLSAR放在所构建的环境中用A3C进行训练. 训练结果表明, 强化学习成功地指导SAR个体选择阶段, 且RLSAR实现了在复杂的环境下对UAV的航线进行规划.

本文还利用训练好的RLSAR模型对不同威胁区下RLSAR、SAR、DE和SSA算法进行路径规划, 3个动作(局部搜索、全局搜索和路径调整)对RLSAR的贡献以及对RLSAR中各个体执行的操作序列进行分析. 所有实验均表明, RLSAR性能明显优于标准SAR与其他比较算法.

本文主要有以下现实意义.

1) RLSAR将优化算法个体作为强化学习中的训练对象, 而非无人机本身, 使得训练出的模型可应用于随机生成的环境, 具备较强的泛化能力.

2) RLSAR将群智能优化算法的算子作为强化学习中的动作仅是一种结合强化学习的思路. 其他类型的群智能优化算法可根据算法设计特点, 针对性地与强化学习相结合, 以增强群智能算法在更新个体时的自适应性.

3) 路径调整在RLSAR中起到至关重要的作用, 这表明针对特定问题设计的局部搜索方法可被结合到群智能优化算法中以提升性能.

4) RLSAR的个体数量仅为8个, 这得益于强化学习对SAR搜索的有效提升. 所提出算法可在高维空间中使用更少的个体优化出更优的路径.

如2)所述, 本文利用强化学习指导SAR的个体自适应地选择算子, 仅是对群智能优化算法的一种改进方式, 该方式引出了群智能优化算法或进化算法在算子设计上可考虑自适应选择的问题. 但是, 强化学习改进群智能优化算法的方式并不局限于所提出方法, 未来将继续探究强化学习与不同类型群智能优化算法的结合策略并应用于解决更多的优化问题.

参考文献(References)

[1] Zhang S Y, Zhang X B, Yuan J, et al. A survey on coverage and exploration path planning with multi-rotor micro aerial vehicles[J]. Control and Decision, 2022, 37(3): 513-529.

[2] Liu F, Pu Z H, Zhang S C. UAV multi-target tracking algorithm based on attention feature fusion[J]. Control and Decision, 2023, 38(2): 345-353.

[3] Li L Y, Mu X H, Chianucci F, et al. Ultrahigh-resolution boreal forest canopy mapping: Combining UAV imagery and photogrammetric point clouds in a deep-learning-based approach[J]. International Journal of Applied Earth Observation and Geoinformation, 2022, 107: 102686.

[4] Bailon-Ruiz R, Bit-Monnot A, Lacroix S. Real-time wildfire monitoring with a fleet of UAVs[J]. Robotics and Autonomous Systems, 2022, 152: 104071.

[5] He Z B, Liu C G, Chu X M, et al. Dynamic anti-collision A-star algorithm for multi-ship encounter situations[J]. Applied Ocean Research, 2022, 118: 102995.

[6] Chao N, Liu Y K, Xia H, et al. A sampling-based method with virtual reality technology to provide minimum dose path navigation for occupational workers in nuclear facilities[J]. Progress in Nuclear Energy, 2017, 100: 22-32.

[7] Wang J K, Li B P, Meng M Q H. Kinematic constrained bi-directional RRT with efficient branch pruning for robot path planning[J]. Expert Systems with Applications, 2021,170: 114541.

[8] Booker L B, Goldberg D E, Holland J H. Classifier systems and genetic algorithms[J]. Artificial Intelligence, 1989, 40(1/2/3): 235-282.

[9] Kennedy J, Eberhart R. Particle swarm optimization[C]. Proceedings of ICNN'95-International Conference on Neural Networks. Perth, 1995: 1942-1948.

[10] Pan W T. A new fruit fly optimization algorithm: Taking the financial distress model as an example[J]. Knowledge-Based Systems, 2012, 26: 69-74.

[11] Mirjalili S, Mirjalili S M, Lewis A. Grey wolf optimizer[J]. Advances in Engineering Software, 2014, 69: 46-61.

[12] Liu Y, Zhang X J, Zhang Y, et al. Collision free 4D path planning for multiple UAVs based on spatial refined voting mechanism and PSO approach[J]. Chinese Journal of Aeronautics, 2019, 32(6): 1504-1519.

[13] Pehlivanoglu Y V, Pehlivanoglu P. An enhanced genetic algorithm for path planning of autonomous UAV in target coverage problems[J]. Applied Soft Computing, 2021, 112: 107796.

[14] Qu C Z, Gai W D, Zhang J, et al. A novel hybrid grey wolf optimizer algorithm for unmanned aerial vehicle (UAV) path planning[J]. Knowledge-Based Systems, 2020, 194: 105530.

[15] Zhang X Y, Lu X Y, Jia S M, et al. A novel phase angle-encoded fruit fly optimization algorithm with mutation adaptation mechanism applied to UAV path planning[J]. Applied Soft Computing, 2018, 70: 371-388.

[16] Shabani A, Asgarian B, Salido M, et al. Search and rescue optimization algorithm: A new optimization method for solving constrained engineering optimization problems[J]. Expert Systems with Applications, 2020, 161: 113698.

[17] Kober J, Bagnell J A, Peters J. Reinforcement learning in robotics: A survey[J]. The International Journal of Robotics Research, 2013, 32(11): 1238-1274.

[18] Montague P R. Reinforcement learning: An introduction, by Sutton, R.S. and Barto, A.G.[J]. Trends in Cognitive Sciences, 1999, 3(9): 360.

[19] Ai B, Jia M X, Xu H W, et al. Coverage path planning for maritime search and rescue using reinforcement learning[J]. Ocean Engineering, 2021, 241: 110098.

[20] Gismondi F, Possieri C, Tornambe A. A solution to the path planning problem via algebraic geometry and reinforcement learning[J]. Journal of the Franklin Institute, 2022, 359(2): 1732-1754.

[21] Qu C Z, Gai W D, Zhong M Y, et al. A novel reinforcement learning based grey wolf optimizer algorithm for unmanned aerial vehicles (UAVs) path planning[J]. Applied Soft Computing, 2020, 89: 106099.

[22] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning[C]. Proceedings of the 33rd International Conference on International Conference on Machine Learning. New York, 2016: 1928-1937.

[23] Cao J Q, Zhang G Y, Xu P. A* initialized mutable gray wolf optimizer for UAV path planning[J]. Computer Engineering and Applications, 2022, 58(4): 275-282.

[24] Storn R, Price K. Differential evolution — A simple and efficient heuristic for global optimization over continuous spaces[J]. Journal of Global Optimization, 1997, 11(4): 341-359.

[25] Jain M, Singh V, Rani A. A novel nature-inspired algorithm for optimization: Squirrel search algorithm[J]. Swarm and Evolutionary Computation, 2019, 44: 148-175.

[26] Tayebi S, Momani S, Arqub O A. The cubic B-spline interpolation method for numerical point solutions of conformable boundary value problems[J]. Alexandria Engineering Journal, 2022, 61(2): 1519-1528.

作者简介

周文娟 (1993–), 女, 硕士生, 从事人工智能及其应用、强化学习等研究, E-mail: wjz0923@foxmail.com;

张超群 (1974–), 女, 副教授, 博士, 从事智能计算、大数据技术与应用等研究, E-mail: 25713893@qq.com;

汤卫东 (1968–), 男, 教授, 博士, 从事形式化方法、云计算与大数据等研究, E-mail: 898402726@qq.com;

易云恒 (1998–), 男, 硕士生, 从事人工智能及其应用等研究, E-mail: 1263014397@qq.com;

刘文武 (1996–), 男, 硕士生, 从事人工智能及其应用等研究, E-mail: 865088716@qq.com;

秦唯栋 (1996–), 男, 硕士生, 从事人工智能及其应用等研究, E-mail: 297295560@qq.com.