计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200300028

基于近端策略优化的 RFID 室内定位算法

李　丽　　郑嘉利　　罗文聪　　全艺璇
广西大学计算机与电子信息学院　南宁 530004
广西多媒体通信与网络技术重点实验室　南宁 530004
(1114235262@qq.com)

摘　要　针对在动态射频识别(Radio Frequency Identification,RFID)室内定位环境中,传统的室内定位模型会随着定位目标数量的增加而导致定位误差增大、计算复杂度上升的问题,文中提出了一种基于近端策略优化(Proximal Policy Optimization,PPO)的 RFID 室内定位算法。该算法将室内定位过程看作马尔可夫决策过程,首先将动作评价与随机动作相结合,然后进一步最大化动作回报值,最后选择最优坐标值。其同时引入剪切概率比,首先将动作限制在一定范围内,交替使用采样后与采样前的新旧动作,然后使用随机梯度对多个时期的动作策略进行小批量更新,并使用评价网络对动作进行评估,最后通过训练得到 PPO 定位模型。该算法在有效减少定位误差、提高定位效率的同时,具备更快的收敛速度,特别是在处理大量定位目标时,可大大降低计算复杂度。实验结果表明,本文提出的算法与其他的 RFID 室内定位算法(如 Twin Delayed Deep Deterministic Policy Gradient(TD3),Deep Deterministic Policy Gradient(DDPG),Actor Critic using Kronecker-Factored Trust Region(ACKTR))相比,定位平均误差分别下降了 36.361%,30.696%,28.167%,定位稳定性分别提高了 46.691%,34.926%,16.911%,计算复杂度分别降低了 84.782%7,70.213%,63.158%。

关键词:RFID;室内定位;深度强化学习;剪切概率比
中图法分类号　TP301.6

RFID Indoor Positioning Algorithm Based on Proximal Policy Optimization

LI Li,ZHENG Jia-li,LUO Wen-cong and QUAN Yi-xuan
School of Computer,Electronics and Information,Guangxi University,Nanning 530004,China
Guangxi Key Laboratory of Multimedia Communications and Network Technology,Nanning 530004,China

Abstract　In the Radio Frequency Identification(RFID) dynamic indoor positioning environment,the positioning error and the computing complexity of traditional indoor positioning model will increase with the increase of the number of positioning targets. This paper proposes an RFID positioning algorithm based on Proximal Policy Optimization(PPO),which regards the positioning as Markov decision-making process. Firstly,the action evalution is combined with random action and the return of the action is then maximized to select the best coordinate value. Meanwhile,under the premise of limiting the action to a certain range,the algorithm introduces clipped probability ratios,using post-sample and pre-sample action alternatesly,then,with stochastic gradient ascent updates multiple epochs policy of minibatch and with the critic network evaluate the action. Finally,the PPO positioning model is obtained by training. This method can effectively reduce the positioning error and improve the positioning efficiency. At the same time,it has a faster convergence speed,especially when dealing with a large number of positioning targets,it can greatly reduce the computational complexity. Experiment results show that,compared with other RFID indoor positioning algorithms,such as Twin Delayed Deep Deterministic policy gradient(TD3),Deep Deterministic Policy Gradient(DDPG) and actor-critic using Kronecker-Factored Trust Region(ACKTR),the average positioning error of the proposed method decreases respectively by 36.361%,30.696% and 28.167%,the positioning stability improves by 46.691%,34.926% and 16.911%,and the computing complexity decreases respectively by 84.782%,70.213% and 63.158%.

Keywords　RFID,Indoor positioning,Deep reinforcement learning,Clipped probability ratios

到稿日期:2020-03-04　返修日期:2020-06-29　　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61761004);广西自然科学基金(2019GXNSFAA245045)
This work was supported by the National Natural Science Foundation of China(61761004) and Natural Science Foundation of Guangxi Province,China(2019GXNSFAA245045).
通信作者:郑嘉利(zjl@gxu.edu.cn)

1　引言

由于 RFID 具有非视距、体积小、成本低、抗干扰能力强、识别速度快且能同时识别多个目标等优点，其在室内定位、产品识别等方面得到广泛应用[1-3]。传统的室内定位方法采用普通的路径传播模型，需要考虑信号衰减和角度等因素[2-4]。然而，基于机器学习的指纹识别方法并不依赖于传播模型，该方法在不同的位置收集数据特征，并为每个点构建指纹[5-7]。例如，文献[8]将 RFID 与计算机视觉相结合，将 RFID 标签粘贴在定位目标上，利用支持向量机和人工神经网络拟合出定位目标的位置，再利用视觉系统检测出定位目标的 3D 位置，并结合 K-means 方法来修正位置精度。文献[9]提出了一种基于支持向量回归(Support Vector Regression，SVR)的机器学习模型，用于对固定目标的定位。该模型在离线阶段学习 RSSI 指纹，因此不需要参考标签在线学习标签位置。

上述方法针对大量的定位目标，提出的浅层神经网络不能很好地拟合出精确的位置。因此，人们将深度学习引入室内定位中，文献[10]提出了一种基于深度神经网络的室内定位算法，该算法首先利用一组自动编码器来逐层对权值进行预训练，然后利用 softmax 函数来确定接收机位置在参考点上的概率，从而估计接收机的位置。文献[11]采用 4 层深度神经网络结构进行定位，通过堆叠去噪自编码器来对网络结构进行预训练，在大量有噪声的样本中，自动学习有效特征，拟合出定位目标的位置。

基于深度学习的方法需要大量的样本训练数据，对硬件设备和数据量的要求较高，随着深度学习的不断发展，深度学习结合强化学习，即深度强化学习[12]逐渐成为主流研究方向。如 Lillicrap 等[13]提出的深度确定性策略梯度算法(Deep Deterministic Policy Gradient，DDPG)，该算法将动作评价方法应用到深度确定性策略梯度中，由动作网络执行确定性动作，评价网络评价动作值。但是，该算法计算复杂度较高，训练时间较长。Yu 等[14]提出使用克罗内克因子信赖域的动作评价算法(Actor Critic using Kronecker-Factored Trust Region，ACKTR)，其使用克罗内克因子来计算自然梯度，并用 KL 散度作为参数更新前后的度量，然后用自然梯度来更新动作和评价网络。该算法不仅可以降低计算复杂度，还可以减少运行时间，但是该算法的方差过高，导致模型在训练时容易出现过拟合。因此 Scott 等[15]提出双延迟深度确定性策略梯度算法(Twin Delayed Deep Deterministic Policy Gradient，TD3)，其使用两个评价网络来对动作值进行评估，取最小评估值，同时对动作网络进行剪切和延迟处理，以防止过高方差。该算法还在 Open AI gym 游戏中表现出优异的性能。

上述算法只是在游戏领域中取得了较好的结果，在工业界的应用中，如在室内定位中，还处于发展阶段。因此本文创新性地将深度强化学习应用到 RFID 室内定位中。将深度强化学习中的动作评价与随机动作相结合，结合剪切概率比，建立近端策略优化算法，并将其用于 RFID 室内定位中。在 RFID 室内定位中，该算法以动作选择最优坐标为目标，使用信任区域来最大化动作，并利用随机梯度方法更新动作，不断优化更新动作，最终动态规划出 RFID 定位目标的具体位置。该方法不仅极大地降低了计算复杂度，还可以防止高方差和高偏差。

2　基于近端策略优化的 RFID 室内定位算法

传统的室内定位方法在面对大量 RFID 室内定位目标时，往往由于各种环境因素的干扰，会产生较大的定位误差。在深度强化学习中，近端策略优化算法考虑了新动作策略和旧动作策略，在新旧策略中加入剪切概率比，并设置一个新的目标函数，这一新的目标函数可以限制动作值稳定在近端范围内，使新的动作策略可以参照旧的动作策略来进行更新。该方法解决了传统强化学习算法高方差和高偏差的问题，并可以快速决定模型的优化方向。因此，该算法不仅具有传统强化学习动态决策的优势，还可以将动作限制在近端范围内，快速决定优化方向。若将其用在 RFID 室内定位中，可以快速确定出定位目标的位置，并确定出位置的优化方向，从而优化出最优位置信息。本文将近端策略优化算法与 RFID 室内定位相结合，提出了一种基于近端策略优化的 RFID 室内定位算法。该算法包含定位模型建立、定位模型训练、定位模型更新 3 个阶段。首先建立模型的动作和评价网络，并结合剪切概率比建立模型目标函数，模型的目标是动作选择最优坐标；然后设置动作值和回报值，并进行模型训练；最后定位模型利用自然梯度和自适应 KL 惩罚系数[16]对定位模型不断进行优化更新，最终通过训练得到室内定位模型。

2.1　建立定位模型训练网络

首先建立两个动作网络，动作网络是根据 $RSSI$ 值与定位目标点位置的关系拟合出定位目标坐标点，动作网络中的智能体动态选择最优坐标值。由于 RFID 室内定位是连续的，对动作网络、动作目标函数及评价网络都使用神经网络，两个动作网络的网络参数分别为 δ_1，δ_2，同时设置网络 K 为两个动作网络的加权平均，网络参数为 w，然后利用 K 网络来执行动作策略。这里的 K 网络是临时存在的，执行完动作后就消失，将 K 网络设置为：

$$K=\eta K+(1-\eta)K,0<\eta<1 \tag{1}$$

其中 η 为加权参数。

定义动作网络为：

$$Q_{wi}(s_t,a_t)_{i=1,2}=r_t+\sum_{t=1}^{\infty}E[V(a_{t+1},s_{t+1})] \tag{2}$$

其中，r_t 为回报值，$V(a_{t+1},s_{t+1})$ 表示下一个动作及状态的评价网络。

定义评价网络，网络参数为 v，具体表示为：

$$V(s_t,a_t)=E(r_t+\gamma V(s_t,a_t)) \tag{3}$$

评价网络用于估计动作的优劣，因此需要设置评估过程，将评估过程定义为优势函数 $A(s_t,a_t)$，具体表示为：

$$A(s_t,a_t)=\phi_t+\gamma\phi_{t+1}+\cdots+\gamma^{T-t+1}\phi_{T-1} \tag{4}$$

其中，ϕ_t 定义为估计函数，具体表示为：

$$\phi_t=r_t+\gamma V(s_{t+1},a_{t+1})-V(s_t,a_t) \tag{5}$$

最后设置模型的目标函数。RFID 室内定位的目的是定位出更精确的位置，而使用传统的定位精度较高的算法来实现此目的的复杂度较大，因此本文引入具有剪切概率比的目

标函数，使动作网络中的智能体能获得最大化的折扣回报，具体定义为：

$$L(s_t,a_t;\theta)=E[L^{CLIP}(s_t,a_t;\theta)-c_1L^{VF}(s_t,a_t;\theta)+c_2S[\varphi(a_t,s_t;\theta)]] \tag{6}$$

其中，c_1，c_2 为目标函数参数，$\varphi(a_t,s_t;\theta)$ 为动作策略函数，策略参数为 θ，具体表示为：

$$\varphi(a_t,s_t;\theta)=\tanh(f(RSSI|s_t);\theta) \tag{7}$$

式(6)中的 $L^{CLIP}(s_t,a_t;\theta)$ 为剪切概率目标函数，表示为：

$$L^{CLIP}(s_t,a_t;\theta)=E[\min(\tau_tA(s_t,a_t),clip(\tau_t,\varepsilon,A(s_t,a_t)))] \tag{8}$$

式(7)中 ε 为超参数，τ_t 为概率比，$\tau_t=\dfrac{\varphi(a_t,s_t;\theta)}{\varphi_{\text{old}}(a_t,s_t;\theta)}$，表示新老动作策略对当前状态采取动作的对应的概率之比，$clip(\tau_t,\varepsilon,A(s_t,a_t))$ 为剪切概率比，具体表示为：

$$clip(\tau_t,\varepsilon,A(s_t,a_t))=\begin{cases}\min(\tau_t,1+\varepsilon)A(s_t,a_t), & A(s_t,a_t)\geqslant 0\\ \max(\tau_t,1-\varepsilon)A(s_t,a_t), & A(s_t,a_t)\leqslant 0\end{cases} \tag{9}$$

剪切概率比可以将动作策略限制在$[1-\varepsilon,1+\varepsilon]$的范围内，这一限制会对剪切后的目标和未被剪切的目标进行最小化的约束，τ_t 的值取决于优势函数是正优势函数还是负优势函数。

式(6)中 $L^{VF}(s_t,a_t;\theta)$ 为平方误差损失函数，具体表示为：

$$L^{VF}(s_t,a_t;\theta)=(V(s_{t+1},a_{t+1})-V(s_t,a_t))^2 \tag{10}$$

式(6)中 $S[\varphi(a_t,s_t;\theta)]$ 为交叉熵，表示随机变量的随机性，用于确保智能体有足够的探索，使定位模型训练更稳定，具体表示为：

$$S[\varphi(a_t,s_t;\theta)]=E[-\log(\varphi(a_t,s_t;\theta))] \tag{11}$$

通过对新老动作策略比值进行剪切约束，并对剪切后的概率比与优势函数乘积的最小值进行优化，再结合误差损失函数和交叉熵，来确保动作有足够探索的同时，模型的计算复杂度也降低了。

2.2 定位模型训练

首先，对定位区域的参考定位目标进行数据采集，得到RSSI值数据集，然后定位模型根据RSSI值与定位坐标建立非线性关系，最后拟合出参考定位目标坐标值，设置动作 a_t 为动作策略函数 $\varphi(a_t,s_t;\theta)$，智能体根据当前状态 s_t 和特征输入 $f(RSSI|s_t)$，执行动作 a_t 并选取出最优坐标值，并得到下一个状态 s_{t+1} 和回报值 r_t，动作策略函数表示为：

$$f(RSSI|s_t)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(RSSI-B)}{2\sigma^2}\right) \tag{12}$$

其中，σ 表示在状态 s_t 下获取的RSSI值的标准差，具体表示为：

$$\sigma=\sqrt{\frac{1}{(N-1)}\sum_{i=1}^{N}(RSSI_i-B)^2} \tag{13}$$

其中，$RSSI_i$ 表示第 i 个 $RSSI$ 值，B 表示 $RSSI$ 平均值。

$$B=\frac{1}{N}\sum_{i=1}^{N}RSSI_i \tag{14}$$

式(12)中的回报值 $r_t=r(s_t,a_t,s_{t+1})$，具体表示如下：

$$r(s_t,a_t,s_{t+1})=\begin{cases}\dfrac{1}{d_t}, & \text{if } 0<d_t<\lambda\\ -d_t, & \text{otherwise}\end{cases} \tag{15}$$

式(12)中 d_t 为距离误差，具体表示为：

$$d_t=\sqrt{(x_t-x_1^t)^2+(y_t-y_1^t)^2} \tag{16}$$

其中，(x_t,y_t) 是第 t 个定位目标的估计位置，$(x_1{}^t,y_1{}^t)$ 为第 t 个定位目标的实际位置。阈值 λ 为定位的平均误差，具体表示为：

$$\lambda=\frac{1}{N}\sum_{t=1}^{N}d_t \tag{17}$$

其中，N 为总的定位目标数。

最后，定位模型不断训练，选取每个位置的最优坐标值，并累加回报值，当定位区域中的目标全部被定位完成时，得到总的回报值，总回报值设置为：

$$R=\sum_{t=1}^{M}\gamma^{-1}r_t \tag{18}$$

其中，M 为线程数。

基于近端策略优化的RFID室内定位算法流程图如图1所示。

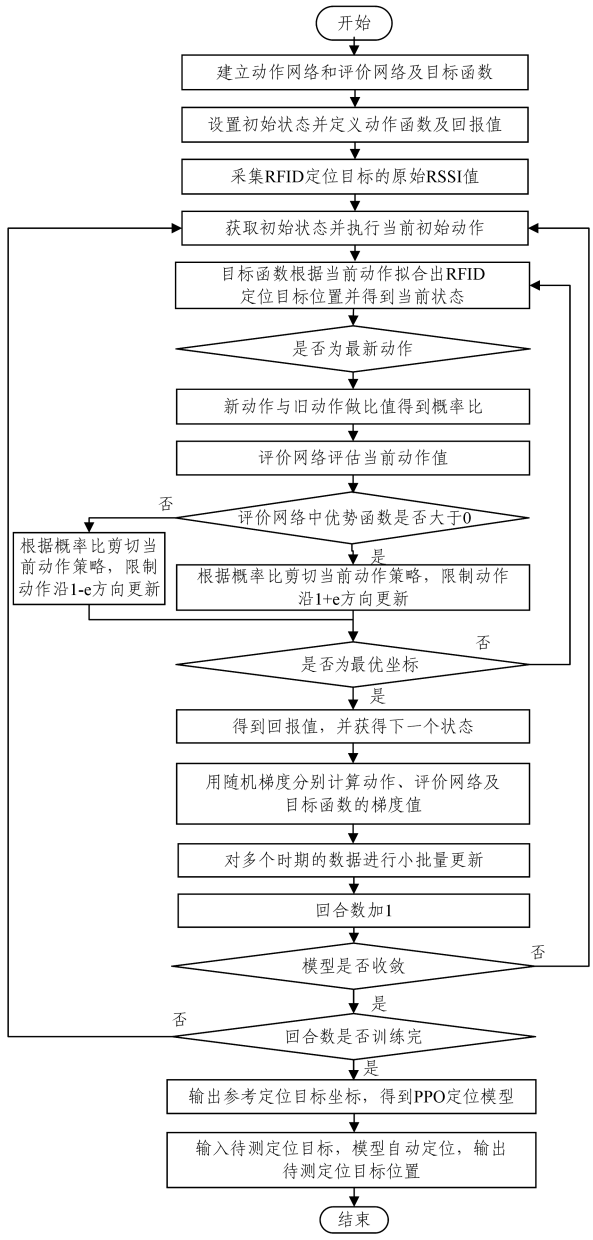


图1 基于PPO的RFID室内定位算法流程图

Fig. 1 Flow chart of RFID indoor positioning algorithm based on PPO

2.3　定位模型更新

由于标准的神经网络更新是对每个数据样本执行梯度更新，若用在 RFID 定位模型中，更新时间比较长，定位效率不高。本文将随机梯度与自适应 KL 惩罚系数相结合来更新梯度，这样可支持多个小批量数据同时更新，以减少定位模型的训练时间，降低样本复杂度。首先，动作网络的梯度更新表示为：

$$dw=\frac{1}{PT}\sum_{p\in P}\sum_{t=0}^{T}E[w+\nabla_{\omega}\log\varphi(a_t,s_t;\theta)A(s_t,a_t)] \tag{19}$$

其中，T 为总的时间步数，P 为轨迹函数，P 表示为：

$$P=\prod_{t=0}^{T}\varphi(a_t|s_t;\theta)p(s_{t+1}|s_t,a_t) \tag{20}$$

其中，$p(s_{t+1}|s_t,a_t)$定义为满足高斯分布的概率函数，$p(s_{t+1}|s_t,a_t)\in N(0,1)$。

然后，评价网络的梯度更新表示为：

$$dv=\arg\min\frac{1}{PT}\sum_{p\in P}\sum_{t=0}^{T}(V(s_t,a_t)-R)^2 \tag{21}$$

最后，目标函数梯度更新表示为：

$$\nabla L(s_t,a_t;\theta)=\arg\max\frac{1}{PT}\sum_{p\in P}\sum_{t=0}^{T}(\nabla_{\theta}L^{CLIP}(s_t,a_t;\theta)-c_1\nabla_{\theta}L^{VF}(s_t,a_t;\theta)+c_2\nabla_{\theta}S[\varphi(a_t,s_t;\theta)]) \tag{22}$$

其中，$\nabla_{\theta}L^{CLIP}(s_t,a_t;\theta)$为剪切概率目标的梯度，表示为：

$$\nabla_{\theta}L^{CLIP}(s_t,a_t;\theta)=\arg\max\frac{1}{PT}\sum_{p\in P}\sum_{t=0}^{T}\nabla_{\theta}\min(\tau_tA(s_t,a_t),g(\tau_t,\varepsilon,A(s_t,a_t))) \tag{23}$$

其中，$\nabla_{\theta}L^{VF}(s_t,a_t;\theta)$为平方误差损失函数的梯度，表示为：

$$\nabla_{\theta}L^{VF}(s_t,a_t;\theta)=\frac{1}{PT}\sum_{p\in P}\sum_{t=0}^{T}\nabla_{\theta}(V(s_{t+1},a_{t+1})-V(s_t,a_t))^2 \tag{24}$$

式(22)中$\nabla_{\theta}S[\varphi(a_t,s_t;\theta)$为交叉熵梯度，表示为：

$$\nabla_{\theta}S[\varphi(a_t,s_t;\theta)=-\frac{1}{PT}\sum_{p\in P}\sum_{t=0}^{T}\nabla_{\theta}\log(\varphi(a_t,s_t;\theta)) \tag{25}$$

用随机梯度分别计算动作网络、评价网络和目标函数的梯度，分别对多个时期的数据进行小批量更新。当动作和评价网络计算完梯度时，再进行反向传播并更新定位模型。当模型进行收敛时，定位误差逐渐趋于稳定。当回合数训练完成时，输出参考定位目标的具体位置。

2.4　定位模型参数设置

本文算法使用 tensorflow[17-18] 架构来实现整个定位模型，动作网络和评价网络均使用多层神经网络[19]，多层神经网络包含两隐藏层，每个隐藏层包含 64 个单元，非线性激活函数为 tanh[20]。模型的参数设置如表 1 所列。

表 1　PPO 定位模型参数

Table 1　Parameters of PPO localization model

超参数	取值
线程数	3
折扣因子	0.99
优化器	Adam
最小分批数	32×8
剪切因子	$\alpha\times0.1$
剪切参数 c_1	1
剪切概率参数 c_2	0.01
单步回报值范围	0.1～1.0
Adam 步数	$2.5\times10^{-4}\times\alpha$
回合数	50 万

由于信号强度波动及噪声的影响，大多数研究都对 *RSSI* 值进行了预处理。本文直接设置动作策略函数为高斯函数，选择最优坐标值。定位模型的环境、动作函数、状态、回报值设置如下。

定位环境的设置。环境由一组位置表示，这些位置由坐标表示，每个坐标还与部署的 RFID 中的 *RSSI* 值集合相关联。具体可以表示为：$RSSI_i^r\rightarrow Locaiton_i$，这里 $RSSI_i^r$ 表示第 i 个定位目标 RSSI 值的集合，r 表示读写器接收到 RSSI 值的时间。智能体通过每次定位出的位置来观察环境，进而了解到周围的 *RSSI* 值。

智能体设置。将智能体设为一个动作函数，随着时间的推移，智能体与环境交互，动作在环境中选择一个随机状态。动作可以根据其是否接近正确的位置点来判断获取回报还是受到惩罚。智能体的目标是通过不同方向的移动探索来近似地从环境中定位出坐标值，若 $position_{present}>position_{best}$，则 $position_{best}=position_{present}$，将当前值复制给最优值，且当前动作选择最优坐标值。

状态设置。状态是智能体可以观察到的一组值的集合，这里我们设定状态元组为：(*RSSI* 值，当前位置(由坐标表示)，目标位置，误差距离)。

回报函数。将回报函数设定为距离误差值的倒数，如果估计定位目标位置与实际定位目标位置之间的距离小于设定的阈值 λ，那么会获得回报，智能体越靠近定位目标，误差值越小，则回报值越大。相反地，若智能体偏离目标或者误差大于阈值 λ，则会受到惩罚。累积回报值越大，定位模型的定位精度越优。

3　实验结果及分析

3.1　数据收集

在本次实验中，定位环境为一个室内教室；实验使用的 RFID 读写器类型为 ALR-9900＋[21]，其主要工作频率为 920～925 MHz，最大增益为 31.6 dBm；实验使用的天线类型为 ALR-9611-CR[22]，增益为 6 dBm。我们将实验部署在一个 $45\times45\,m^2$ 的室内环境中，分别将 6 个定位天线装在 6 个 3 m 高的天线架上，并放置在墙的角落。将区域划分为均匀大小的网格，将参考定位目标放在网格中，读写器以 0.4 的采样率对参考目标的数据进行读取，针对每个参考目标测试 10 组数据，采用 m 次测量法对参考目标的 *RSSI* 值进行累加处理，并将数据作为输入数据。

3.2　实验评估

3.2.1　定位目标静止与移动时的影响

为了检验定位模型针对移动定位目标的定位性能，比较了定位目标移动时与静止时的定位结果，分别获取了参考定位目标静止时与移动时的 *RSSI* 值，并分别比较了模型训练和模型定位时的结果。

(1)模型训练

图 2 给出了定位目标移动时和静止时的收敛速度。在定位目标静止的情况下，在定位模型训练前期，定位目标移动时的回报值比静止时的回报值低，但是在训练后期，两种情况的

回报值基本相同，且最后模型都收敛了。

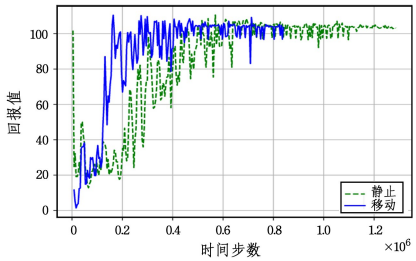


图 2 比较目标移动时和静止时的收敛速度

Fig. 2 Convergence rates comparison between moving and static targets

(2)模型定位

图 3 和图 4 分别给出了定位目标在移动时和静止时模型在定位阶段回报值的最大值、平均值和最小值。如图 3 和图 4 所示，定位模型的最大值和平均值均在 100 以上，这表明定位模型可以定位出目标点位置。从图 3 可以看出，当定位目标移动时，其最大值、最小值及平均值都有一定程度的波动，但均稳定在 100 以上，且最大回报值还能达到 140 以上，这表明当定位目标移动时，智能体的探索度较高。如图 4 所示，当定位目标静止时，其最大值、最小值及平均值可以达到 104 及以上。综上，当定位目标移动时，定位模型也可以定位出目标位置，具有较好的探索能力。

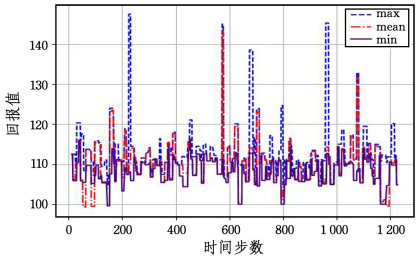


图 3 目标移动时的定位能力

Fig. 3 Positioning ability when target is moving

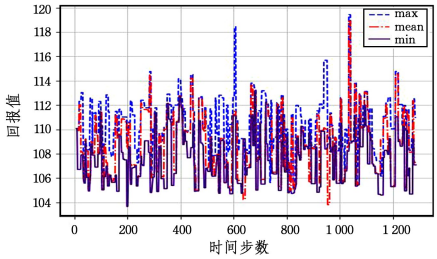


图 4 目标静止时的定位能力

Fig. 4 Positioning ability when target is static

表 2 列出了定位目标在移动时和静止时的定位精度最大值 max、最小值 min、平均值 mean、标准差 std。

表 2 比较定位目标移动和静止时的定位精度

Table 2 Positioning accuracy comparison between moving and static targets

(单位：m)

状态	max	min	mean	std
移动	2.13474	0.33822	0.62497	0.22006
静止	1.97049	0.27209	0.58361	0.06879

由表 2 可知，当定位目标移动时，最大定位误差为 2.13474 m，最小定位误差为 0.3382 m；当其静止时，最大定位误差为 1.9704 m，最小定位误差为 0.27209 m。在定位目标移动时，定位误差会略微增加。综上所述，PPO 在定位目标移动时，也可以动态地适应环境，具有较好的鲁棒性和环境适应性。

3.3.2 与其他算法比较

为了测试 PPO 室内定位模型的性能，PPO 室内定位算法还分别从模型的收敛速度、定位能力、定位精度、模型稳定性、计算复杂度、定位效率等方面与 ACKTR 定位、TD3 定位、DDPG 定位算法进行比较。

(1)模型训练

图 5 给出了各个定位模型在训练阶段的收敛速度。DDPG 的收敛速度最慢，在训练前期，该模型的回报值不断增加，而当训练步数达到 50 万步时，其回报值突然下降，经过一段时间后，其回报值才继续增加，直到模型收敛。PPO，ACKTR，TD3 则可以不断学习直至最后模型收敛。但是 PPO 可以在短时间内收敛，表明 PPO 定位模型在收敛速度上优于其他几种算法。

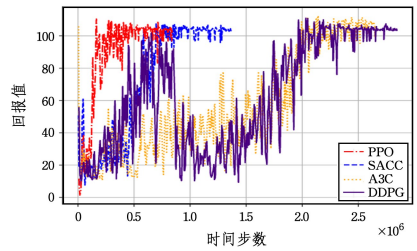


图 5 比较 4 种算法的收敛速度

Fig. 5 Compare convergence rates of 4 positioning algorithms

(2)模型定位

1)定位能力

图 6—图 9 给出了各个定位模型在实际定位阶段的定位能力，这里用回报值来表示。

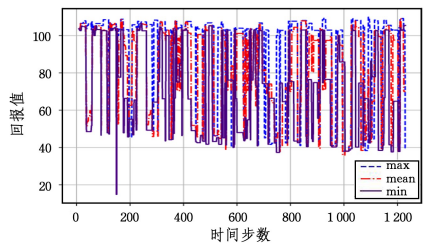


图 6 DDPG 模型的定位能力

Fig. 6 Positioning ability of DDPG model

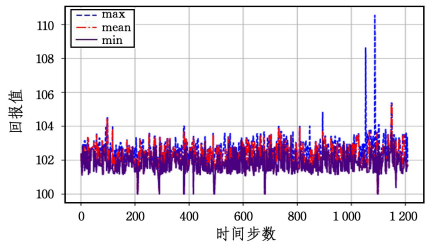


图 7 TD3 模型的定位能力

Fig. 7 Positioning ability of TD3 model

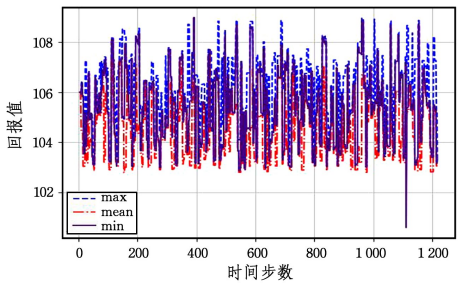


图 8　ACKTR 模型的定位能力

Fig. 8　Positioning ability of ACKTR model

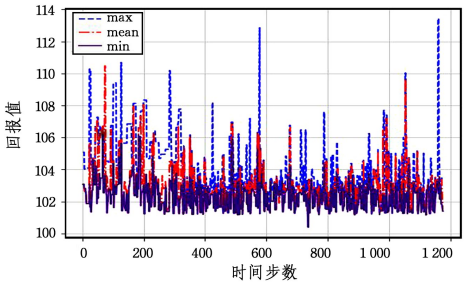


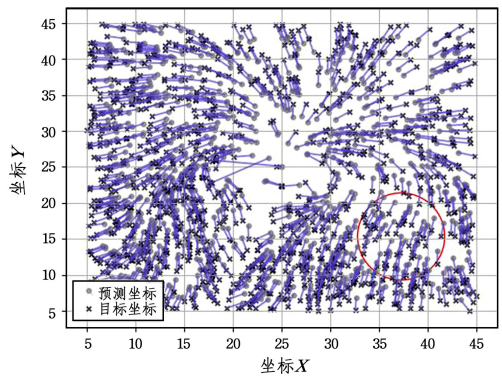
图 9　PPO 模型的定位能力

Fig. 9　Positioning ability of PPO model

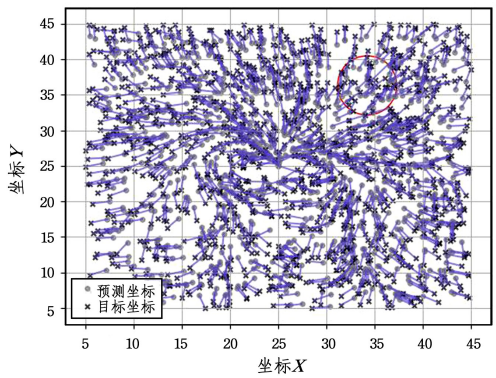
从图中可以看出，DDPG 的定位能力最弱，在前 100 个时间步数中，DDPG 还可以较为精确地定位出目标点位置，但当超过 100 个时间步数后，其定位能力减弱，且回报值波动幅度较大，只能粗略定位出目标点位置。TD3 和 ACKTR 的定位能力比 DDPG 好，两者的平均回报值基本在 100 以上。TD3 和 ACKTR 都可以较为精确地定位出目标点位置。PPO 的定位能力最好，其最大回报值、最小回报值和平均回报值均在 103 以上，并且 PPO 在每次定位过程中都可以迅速精确地定位出目标点位置。这表明模型先拟合出坐标值，再选出最优坐标的策略可以提高模型的定位能力。

2）定位精度

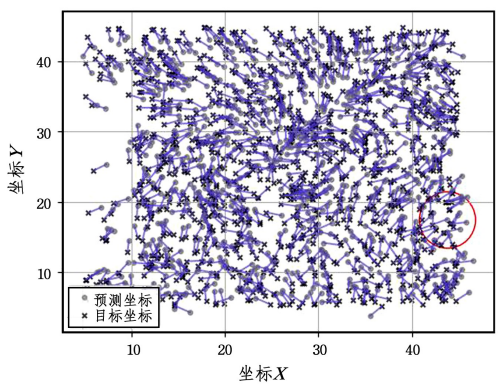
在图 10 中，圆圈表示实际坐标值与预测坐标值的最大误差，圆圈越大，最大定位误差越大，“×”代表目标定位点，“●”代表预测定位点，连线代表目标定位点与预测定位点之间的距离。从图 10 和表 3 可以看出，PPO 定位能力最好，与 DDPG 相比，其平均误差下降了 36.361%，与 TD3 相比，其平均误差下降了 30.696%，与 ACKTR 相比，其平均误差下降了28.167%。这是由于 PPO 重新设计了目标函数，先拟合出坐标值，同时利用剪切概率比，将动作限制在一定范围内，然后利用动作评价方法选择最优坐标值。该方法可以有效减少噪声等对定位过程的干扰，提高了定位的精确性。而 ACKTR，TD3 和 DDPG 只是传统的奖励回报机制，它们在学习过程中还是通过选择最优 *RSSI* 值来拟合出定位目标，其自我学习能力和适应环境性能相比 PPO 较差。由表 3 可知，DDPG，TD3，ACKTR，PPO 的最大定位误差分别为 2.521 8 m，2.457 8 m，2.117 8 m，1.570 4 m。出现最大定位误差的原因是当定位目标在边缘或角落时，由于距离和角度原因，信号强度减弱，定位误差变大，然而 PPO 在信号强度较弱的情况下也可以较为精确地定位出目标点位置。



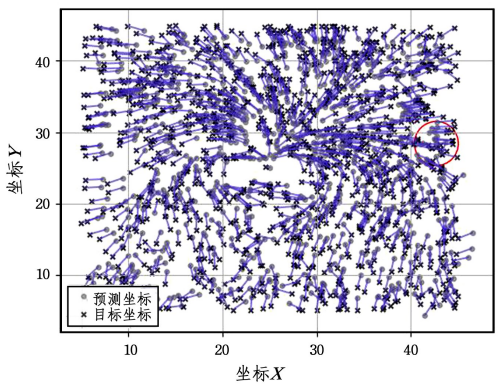
(a) DDPG 模型的定位精度



(b) TD3 模型的定位精度



(c) ACKTR 模型的定位精度



(d) PPO 模型的定位精度

图 10　比较 4 种算法的定位精度

Fig. 10　Compare positioning accuracy of 4 algorithms

表 3 比较 4 种算法的定位误差

Table 3 Compare positioning error of 4 algorithms

(单位:m)

算法	max	min	mean	std
PPO	1.5704	0.2720	0.58366	0.1088
ACKTR	2.1782	0.3369	0.81253	0.1272
TD3	2.4578	0.5964	0.84218	0.1468
DDPG	2.5218	0.5122	0.91715	0.1596

3)模型稳定性

由图 6—图 10 和表 3 可知,DDPG 定位模型的稳定性最差,TD3 和 ACKTR 次之,PPO 的稳定性最好。首先,DDPG 的收敛速度最慢,且存在波动幅度较大的情况,TD3 和 ACKTR 次之,PPO 则可以快速收敛。然后在定位阶段,DDPG 的平均回报值最低,回报值波动幅度很大,在定位过程中不能很好地拟合出定位目标位置,定位误差值较大;而 PPO 可以直接找出最优坐标值,可以随着环境变化不断调整网络,动态定位出目标位置。最后,PPO 的标准差均小于其他算法,与 ACKTR,TD3,DDPG 相比,其稳定性分别提高了 16.911%,34.926%,46.691%。标准差越小,定位模型定位波动性越小,模型稳定性越好。这是由于定位目标越多,反射的信号就越多,PPO 就可以随着定位环境不断调整网络,并快速地选取最优坐标值。

4)时间复杂度

表 4 列出了 4 种算法的模型训练时间和单个定位目标的定位时间。由表可知,PPO 的训练时间和定位时间最短,比 ACKTR 减少了 63.158%,比 TD3 减少了 70.213%,比 DDPG 减少了 84.782%。这是因为 PPO 的新的目标函数和剪切概率比先拟合出定位目标位置并将其限制在一定范围内,再动态选出最优坐标值,同时结合随机梯度小批量更新定位模型,所以缩短了模型的训练时间。在实际定位过程中,PPO 会快速地直接定位出定位坐标,而在 ACKTR 中,智能体还需探索策略,在 TD3 和 DDPG 中,智能体除了需要探索策略外,在到达目标点时还需绕着目标点转圈,才能定位出目标点。

表 4 比较 4 种算法的计算复杂度

Table 4 Compare computing complexity of 4 positioning algorithms

算法	训练时间/min	测试时间/s
PPO	14	0.52
ACKTR	38	2.12
TD3	47	2.47
DDPG	92	4.31

5)定位效率

为了比较定位模型的定位效率,设置定位目标个数为 45000 个,比较了 4 种算法在定位时间不断增加时,能准确定位出的目标个数。由图 11 可知,与 TD3,DDPG,ACKTR 这几种算法相比,PPO 定位模型在单位时间内能够定位的目标个数较多。随着定位时间的不断增加,PPO 定位模型可以逐渐定位出所有目标,最终能完成全部定位,定位效率较高。而其他几种算法在当定位目标个数到达一定值时,随着时间的增加,定位出的目标个数基本不变。TD3 和 DDPG 的定位效率较差,这是因为 TD3 和 DDPG 是单线程网络,在面对大量数据时,泛化能力有限,所以定位效率较低。

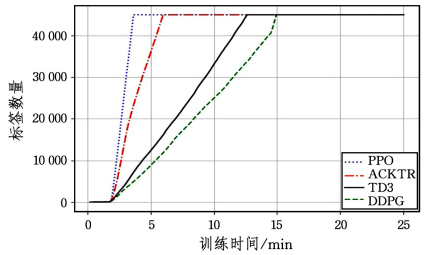


图 11 比较 4 种算法的定位效率

Fig. 11 Compare positioning efficiency of 4 algorithms

结束语 本文通过对现有的室内定位模型的分析,提出了一种基于近端策略优化的 RFID 室内定位算法。该算法首先将动作评价与随机动作相结合,选择最优坐标值,同时引入剪切概率比,将动作限制在一定范围内,交替使用采样后与采样前的新旧动作,然后使用随机梯度对多个时期的动作策略进行小批量更新,并用评价网络对动作进行评估,最后训练得到 RFID 室内定位模型。实验结果表明,本算法与传统基于强化学习的室内定位算法相比,在收敛速度、定位精度和稳定性、定位效率、计算复杂度等方面都有较大的提升。

本文的后续工作将着眼于 RFID 室内定位模型结合实际的定位场景,引入基于二阶梯度的近端策略优化来加快 RFID 定位模型的训练速度,通过多线程动作评价并行训练,来动态适应实际环境,进一步提高定位的准确度。

参 考 文 献

[1] FENG Z, KAISER T. Localization with RFID[M]. New York: John Wiley & Sons, Ltd., 2016: 220-248.

[2] CHOI J S, LEE H, ENGELS D W, et al. Passive UHF RFID-Based Localization Using Detection of Tag Interference on Smart Shelf [J]. IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), 2012, 42(2): 268-275.

[3] MUGAHID O, YUNT G. Indoor distance estimation for passive UHF RFID tag based on RSSI and RCS [J]. Measurement, 2018, 127(10): 425-430.

[4] METTES P, GEMERT J C V, SNOEK C G M. Spot On: Action Localization from Pointly-Supervised Proposals[C]// European Conference on Computer Vision. 2016: 437-453.

[5] HAN K, CHO S H. Advanced LANDMARC with adaptive k-nearest algorithm for RFID location system[C]// 2010 2nd IEEE International Conference on Network Infrastructure and Digital Content. Beijing, China: IEEE, 2010: 595-598.

[6] CHAN M, ZHANG X. Experiments for Leveled RFID Localization for Indoor Stationary Objects[C]// 2014-11th International Conference on Information Technology: New Generations (ICITNG'14). Las Vegas, NV, USA: IEEE, 2014: 1-7.

[7] ZHAO Y, LIU K, MA Y, et al. Similarity Analysis-Based Indoor

Localization Algorithm with Backscatter Information of Passive UHF RFID Tags[J]. IEEE Sensors Journal, 2016, 17(99): 1-9.

[8] BERZ E L, TESCH D A, HESSEL F P. Machine-learning-based system for multi-sensor 3D localization of stationary objects[J]. IET Cyber-Physical Systems: Theory & Applications, 2018, 3(2): 81-88.

[9] JAEHYUN Y, KIM H. Target Localization in Wireless Sensor Networks Using Online Semi-Supervised Support Vector Regression[J]. Sensors, 2015, 15(6): 12539-12559.

[10] WU G S, TSENG P H. A Deep Neural Network-Based Indoor Positioning Method using Channel State Information[C]// 2018 International Conference on Computing, Networking and Communications(ICNC). Maui, HI, USA: IEEE Computer Society, 2018.

[11] LIU K, ZHANG W, ZHANG W D, et al. A Wireless Positioning Method Based on Deep Neural Network[J]. Computer Engineering, 2016, 42(7): 82-85.

[12] SUTTON R, BARTO A. Reinforcement Learning: An Introduction(second edition)[M]. Cambridge: MIT Press, 2018: 1-50.

[13] LILLICRAP T, HUNT P, PRITZEL J, et al. Continuous control with deep reinforcement learning[J]. arXiv: 1509.02971, 2015.

[14] YU H W, ELMAN M, SHUN L, et al. Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation[J]. arXiv: 1708.05144, 2017.

[15] SCOTT F, HERKE V H, DAVID M. Addressing Function Approximation Error in Actorcritic methods[J]. arXiv: 1802.09477v3, 2018.

[16] JOHN S, LEVINE S, MORITZ P, et al. Trust Region Policy Optimization[J]. Computer Science, 2015(3): 1889-1897.

[17] MARTÍ N A, ISARD M, MURRAY D G. A computational model for TensorFlow: an introduction[C]// ACM Sigplan International Workshop on Machine Learning and Programming Languages. Barcelona, Spain: ACM, 2017: 1-7.

[18] ABADI M. TensorFlow: learning functions at scale[J]. Acm Sigplan Notices, 2016, 51(9): 1.

[19] ZHAI X, ALI A A S, AMIRA A, et al. MLP Neural Network Based Gas Classification System on Zynq SoC[J]. IEEE Access, 2017, 4(99): 8138-8146.

[20] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. arXiv: 1409.1556, 2014.

[21] ZHAO Y, LIU K, MA Y, et al. Similarity Analysis-Based Indoor Localization Algorithm with Backscatter Information of Passive UHF RFID Tags [J]. IEEE Sensors Journal, 2016, 17(99): 1-9.

[22] MUGAHID O, YUN T G. Indoor distance estimation for passive UHF RFID tag based on RSSI and RCS[J]. Measurement, 2018, 127(10): 425-430.

LI Li, born in 1994, postgraduate. Her main research interests include information processing, communication networks, reinforcement learning and Internet of things.

ZHENG Jia-li, born in 1979, professor. His main research interests include Internet of things, RFID and artificial intelligence.