(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 109212476 B
(45)授权公告日 2023.03.14

(21)申请号 201811085404.X

(22)申请日 2018.09.18

(65)同一申请的已公布的文献号
申请公布号 CN 109212476 A

(43)申请公布日 2019.01.15

(73)专利权人 广西大学
地址 530004 广西壮族自治区南宁市西乡塘区大学东路100号

(72)发明人 郑嘉利 李丽

(51)Int.Cl.
G01S 5/08(2006.01)
H04W 64/00(2009.01)

(56)对比文件
CN 106910351 A,2017.06.30
CN 107247260 A,2017.10.13
WO 2018053187 A1,2018.03.22
CN 107064913 A,2017.08.18
US 2017024643 A1,2017.01.26
CN 108540929 A,2018.09.14
刘侃等.一种基于深度神经网络的无线定位方法.《计算机工程》.2016,第42卷(第07期),全文.
郭宪.基于深度增强学习的智能体行为演进研究综述.《中国新通信》.2017,第19卷(第17期),全文.
温暖等.深度强化学习在变体飞行器自主外形优化中的应用.《宇航学报》.2017,第38卷(第11期),全文.
杨子薇等.基于标签分组的新型Q值防碰撞算法.《计算机科学》.2018,第45卷(第09期),全文.
Yuenan Hou 等.A novel DDPG method with prioritized experience replay.《2017 IEEE International Conference on Systems, Man, and Cybernetics》.2017,全文.
Haibo Shi 等.Model-based DDPG for motor control.《2017 International Conference on Progress in Informatics and Computing (PIC)》.2018,全文.
(续)

审查员 曹萌媛

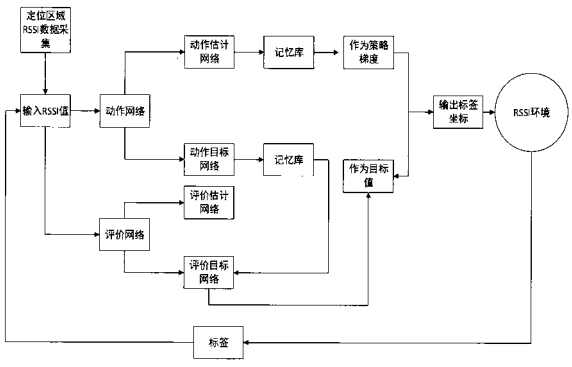
权利要求书2页 说明书5页 附图2页

(54)发明名称

一种基于DDPG的RFID室内定位算法

(57)摘要

本发明涉及无线射频识别(Radio Frequency Identification,RFID)室内定位技术,具体地说是一种基于深度确定性梯度下降(Deep Deterministic Policy Gradient,DDPG)的RFID室内定位算法,包括:建立动作网络和评价网络,其中,动作网络包括动作估计网络和动作目标网络;评价网络包括评价估计网络和评价目标网络。使用动作-评价方法让策略梯度单步更新,同时策略梯度能被用在连续动作上进行筛选,而且在筛选的过程中加入确定性,在连续动作上输出一个动作值,从而确定目标标签的位置。由于RFID室内定位动作是连续的,DDPG与RFID室内定位相结合,很好的解决了定位连续性的问题。本发明与传统的基于神经网络的室内定位算法相比,在定位动作上更连续,进一步提高了定位精度,特别适用于标签信息较庞大的情况。



CN 109212476 B

[转续页]

[接上页]

(56)对比文件

Eduardo Bejar 等.Deep reinforcement learning based neuro-control for a two-dimensional magnetic positioning system.《2018 4th International Conference on Control, Automation and Robotics 》.2018,全文.

翟建伟.基于深度Q网络算法与模型的研究.《中国优秀博硕士学位论文全文数据库(硕士)信息科技辑》.2018,全文.

1.一种基于DDPG的RFID室内定位算法,其特征在于,包括以下步骤:

步骤1)对区域内的M个RFID样本标签的RSSI值进行采集,获得原始训练数据;

步骤2)初始化噪声,利用动作网络的Q估计网络学习,在每个动作中加入噪声,更新状态并获得RFID样本标签最优的RSSI值,将学习到的经验和数据存入记忆库中;

步骤3)训练神经网络:建立动作网络$Q(s,a|\theta^{Q})$和评价网络$\mu(s|\theta^{\mu})$,再分别建立这两个网络的目标网络:$Q'\leftarrow Q$,$\mu'\leftarrow\mu$,目标网络获得下一个状态动作函数,根据评价损失函数更新评价网络,同时根据策略梯度更新动作网络,最后再更新动作网络和评价网络的权重目标网络,使其跟踪学习网络,输出RFID样本标签对应的具体位置,最终得到DDPG定位模型;

步骤4)精准定位:当携带有RFID标签的待定位目标进入检测区域,读写器获取标签信息及RSSI信号强度值,然后将这些数据传至计算机并输入到训练好的DDPG定位模型中,模型准确识别数据并输出待定位目标的具体位置。

2.根据权利要求1所述的一种基于DDPG的RFID室内定位算法,其特征在于,所述步骤2)中初始化噪声,具体包括:初始化噪声分布N,每个动作策略添加一个噪声,执行当前动作a_t并观察当前回报值r_t,然后观察得到下一个状态s_{t+1},在记忆库R中保存经验(s_t,a_t,r_t,s_{t+1}),并获得当前RFID标签的最优RSSI值。

3.根据权利要求1所述的一种基于DDPG的RFID室内定位算法,其特征在于,所述步骤2)中,采用记忆回放的方法,先建立一个记忆库,将部分采样样本收集起来,每次优化是从记忆库中随机取出一部分进行优化,进行小批量的学习,这样可以在不同类型单元的不同任务之间有效学习,减少部分动作不稳定性问题。

4.根据权利要求1所述的一种基于DDPG的RFID室内定位算法,其特征在于,所述步骤2)中的学习过程是一个不断递归的过程,符合贝尔曼方程。

5.根据权利要求1所述的一种基于DDPG的RFID室内定位算法,其特征在于,所述步骤3)中训练神经网络,具体包括:

a)取记忆并训练:从记忆库中取出部分随机样本,表示为(s_i,a_i,r_i,s_{i+1}),然后训练更新目标网络,学习过程可表示为

$y_i=r_i+\gamma Q'(s_{i+1},u'(s_{i+1}|\theta^{u'})|\theta^{Q'})$

其中y_i代表目标网络,r_i代表在i学习过程中的回报值,$\theta^{u'}$、$\theta^{Q'}$代表目标权重,γ代表折扣因子;

b)根据最小损失函数更新评价网络:

$$L=\frac{1}{N}\sum_i(y_i-Q(s_i,a_i|\theta^{Q}))^2$$

其中L代表损失函数;

c)根据策略梯度更新动作网络:

$$\nabla_{\theta^{\mu}}J\cong\frac{1}{N}\sum_i\nabla_a Q(s,a|\theta^{Q})|_{s=s_i,a=\mu(s_i)}\nabla_{\theta_{\mu}}\mu(s|\theta^{\mu})|_{s_i}$$

其中$\nabla_{\theta^{\mu}}J$代表梯度,用动作网络的方法调整权重值;

d)更新目标网络,即权重更新:

评价网络权重更新:$\theta^{Q'}\leftarrow\tau\theta^{Q}+(1-\tau)\theta^{Q'}$;

动作网络权重更新:$\theta^{\mu'}\leftarrow\tau\theta^{\mu}+(1-\tau)\theta^{\mu'}$。

6.根据权利要求1所述的一种基于DDPG的RFID室内定位算法，其特征在于，所述步骤3)中，评价网络类似于策略评估，用于估计动作值函数$\mu(s|\theta^{\mu})$，动作以评价所指导的方向更新策略参数，深度确定性策略梯度DDPG为：

$\nabla_{\theta}J(\mu_{\theta}) = E_{s\sim\rho^{\mu}}\ [\nabla_{\theta}\mu_{\theta}(s)\nabla_{a}Q^{\mu}(s,a)|_{a=\mu_{\theta}(s)}]$。

7.根据权利要求1所述的一种基于DDPG的RFID室内定位算法，其特征在于，所述步骤3)中，策略梯度包括：在连续行动空间使用离线学习算法进行优化，采用e-greedy贪婪策略，以一定的概率使用随机函数，而在剩下的情况下使用最优行动，最终得到一个确定的动作，这个动作不需要从概率分布中采样，相当于当前状态下的最优策略。

8.根据权利要求1所述的一种基于DDPG的RFID室内定位算法，其特征在于，所述步骤3)中，目标网络通过设置一个不会大幅更新的模型，使模型计算的值函数在一定程度上减少波动，使RFID标签定位位置更稳定，通过采用滑动平均的方法更新目标网络：

$\theta_{t+1}\leftarrow\tau\theta_{t}+(1-\tau)\theta'_{t}$

τ设置为非常接近1的数，这样目标网络的参数θ不会发生太大的变化。

一种基于DDPG的RFID室内定位算法

技术领域

[0001] 本发明涉及无线射频识别(Radio Frequency Identification,RFID)中的室内定位技术,具体地说,是一种基于深度确定性策略(Deep Deterministic Policy Gradient,DDPG)的RFID室内定位算法。

背景技术

[0002] 随着通信技术和物联网的发展,智能终端及移动生活的普及,人们的生活和工作中都需要应用到基于位置提供的定位服务,对定位要求也越来越高,所需的定位技术也从室外定位发展到室内定位。室内定位还是室外定位,这是根据定位对象的应用场景确定的。在室外定位中,基于卫星导航的定位技术已经趋于成熟,但是室外定位由于受稠密植被和大部分建筑物的影响,会造成定位不够准确,甚至不能定位。为了满足人们对高精度、低成本定位技术的需求,室内定位成了人们研究的热点。目前,室内定位在人员定位追踪、资产管理、安防救援和商品零售的领域有广泛的应用前景。

[0003] 当前,RFID室内定位技术,有辅助GPS技术、红外线定位技术、超宽带(UHB)定位技术、超声波定位技术、WIFI定位技术、RFID定位技术、蓝牙定位技术、计算机视觉定位技术、图像分析定位技术、光跟踪技术、信标定位技术等等定位技术。其中RFID定位技术是一种比较高效的定位方法,它具有能耗低,实施成本少,测量性高及定位精度高等特点。RFID室内定位基于不同的解决思路主要有四种测距思路,分别是基于信号到达时间(TOA)测距法,基于信号到达时间差(TDOA)测距法,基于信号到达角度(AOA)测距法,基于信号到达强度(Received Signal Strength Indication,RSSI)测距法。本发明主要是基于信号到达强度测距方法。

[0004] 目前,许多人将机器学习的方法用在室内定位方法中,如基于贝叶斯室内定位的分层模型完成无线网络的精确估计,模型在训练时间上提升了许多,引入了完全自适应零捕捉位置方法的概念。基于机器学习指纹的定位算法,可以提供比其他现有的指纹方法更高的定位精度,降低了定位成本,突破了机器学习定位方法只能适用于有源标签的弊端,将范围扩大到了无源标签上。基于粒子波模型的定位算法,需要通过大量粒子群模拟状态分布,然后根据观察结果更新他们的权重模型,粒子通常收敛于最可能的用户位置,收敛成本比较高。

发明内容

[0005] 本发明的目的是提供一种基于DDPG的RFID室内定位算法,利用强化学习中的深度确定性策略建立多种神经网络,通过动作-评价策略确定连续动作的输出,从而构建DDPG定位模型,最终得到RFID待测目标的具体位置。

[0006] 为实现上述目的,本发明提供了如下方案:

[0007] 一种基于DDPG的RFID室内定位算法,包括:

[0008] 步骤1)对区域内的M个RFID样本标签的RSSI值进行采集,获得原始训练数据;

[0009] 步骤2)初始化噪声,利用动作网络的Q估计网络学习,在每个动作中加入噪声,更新状态并获得RFID样本标签最优的RSSI值,将学习到的经验和数据存入记忆库中;

[0010] 步骤3)训练神经网络:建立动作网络$Q(s,a|\theta^{Q})$和评价网络$\mu(s|\theta^{\mu})$,再分别建立这两个网络的目标网络:$Q'\leftarrow Q$,$\mu'\leftarrow\mu$,目标网络获得下一个状态动作函数,根据评价损失函数更新评价网络,同时根据策略梯度更新动作网络,最后再更新动作网络和评价网络的权重目标网络,使其跟踪学习网络,输出RFID样本标签对应的具体位置,最终得到DDPG定位模型;

[0011] 步骤4)精准定位:当携带有RFID标签的待定位目标进入检测区域,读写器获取标签信息及RSSI信号强度值,然后将这些数据传至计算机并输入到训练好的DDPG定位模型中,模型准确识别数据并输出待定位目标的具体位置。

[0012] 作为本发明的进一步改进,所述步骤2)中初始化噪声,具体包括:初始化噪声分布N,构造探索策略μ',每个动作策略添加一个噪声,执行当前动作a_t并观察当前回报值r_t,然后观察得到下一个状态s_{t+1},在记忆库R中保存经验(s_t,a_t,r_t,s_{t+1}),并获得当前标签的最优RSSI值。

[0013] 作为本发明的进一步改进,所述步骤2)中,采用记忆回放的方法,先建立一个记忆库,将部分采样样本收集起来,每次优化是从记忆库中随机取出一部分进行优化,进行小批量的学习,这样可以在不同类型单元的不同任务之间有效学习,减少部分动作不稳定性问题。

[0014] 作为本发明的进一步改进,所述步骤2)中的学习过程是一个不断递归的过程,符合贝尔曼方程。

[0015] 作为本发明的进一步改进,所述步骤3)中训练神经网络,具体包括:

[0016] a)取记忆并训练:从记忆库中取出部分随机样本,表示为(s_i,a_i,r_i,s_{i+1}),然后训练更新目标网络,学习过程可表示为

[0017] $y_i=r_i+\gamma Q'(s_{i+1},u'(s_{i+1}|\theta^{u'})|\theta^{Q'})$

[0018] 其中y_i代表目标网络,r_i代表在i学习过程中的回报值,$\theta^{u'}$、$\theta^{Q'}$代表目标权重,γ代表折扣因子;

[0019] b)根据最小损失函数更新评价网络:

[0020] $$L=\frac{1}{N}\sum_i(y_i-Q(s_i,a_i|\theta^{Q}))^2$$

[0021] 其中L代表损失函数;

[0022] c)根据策略梯度更新动作网络:

[0023] $$\nabla_{\theta^{\mu}}J\cong\frac{1}{N}\sum_i\nabla_a Q(s,a|\theta^{Q})|_{s=s_i,a=\mu(s_i)}\nabla_{\theta_\mu}\mu(s|\theta^{\mu})|_{s_i}$$

[0024] 其中$\nabla_{\theta^{\mu}}J$代表梯度,用动作网络的方法调整权重值;

[0025] d)更新目标网络,即权重更新:

[0026] 评价网络权重更新:$\theta^{Q'}\leftarrow\tau\theta^{Q}+(1-\tau)\theta^{Q'}$;

[0027] 动作网络权重更新:$\theta^{\mu'}\leftarrow\tau\theta^{\mu}+(1-\tau)\theta^{\mu'}$。

[0028] 作为本发明的进一步改进,所述步骤3)中,评价网络类似于策略评估,用于估计动作值函数$\mu(s|\theta^{\mu})$,动作以评价所指导的方向更新策略参数,深度确定性策略梯度DDPG为:

[0029] $\nabla_\theta J(\mu_\theta) = E_{s\sim\rho^\mu}[\nabla_\theta \mu_\theta(s) \nabla_a Q^\mu(s,a)|_{a=\mu_\theta(s)}]$。

[0030] 作为本发明的进一步改进，所述步骤3)中，策略梯度包括：在连续行动空间使用离线学习算法进行优化，采用e-greedy贪婪策略，以一定的概率使用随机函数，而在剩下的情况下使用最优行动，最终得到一个确定的动作，这个动作不需要从概率分布中采样，相当于当前状态下的最优策略。

[0031] 作为本发明的进一步改进，所述步骤3)中，目标网络通过设置一个不会大幅更新的模型，使模型计算的值函数在一定程度上减少波动，使RFID定位位置更稳定，通过采用滑动平均的方法更新目标网络：

[0032] $\theta_{t+1} \leftarrow \tau\theta_t + (1-\tau)\theta'_t$

[0033] τ设置为非常接近1的数，这样目标网络的参数θ不会发生太大的变化。

[0034] 本发明的有益效果为：

[0035] (1)由于RFID室内定位中RSSI值是连续读取的，因此若要筛选出最优RSSI值，这个动作也应当是连续的，利用策略梯度的连续性，动作-评价让策略梯度单步更新，可以很好的解决这个问题。

[0036] (2)利用策略梯度与深度Q网络相结合，包含了深度Q网络的取记忆学习，反向传播，梯度更新，自动探索学习等，解决不同标签输出的RSSI定位问题。

[0037] (3)为了避免确定性策略陷入局部最优化学习的问题，策略中加入部分噪声，使动作探索更广泛和高效。

[0038] (4)建立多个神经网络，包括动作网络和评价网络，动作网络和评价网络分别包含各自的估计和现实网络，利用评价网络指导动作网络，动作网络利用梯度策略不断地修改更新，最终网络快速收敛并选择出最优RSSI值，输出RFID标签具体位置。

[0039] (5)当待测目标进入定位区域时，动作网络从记忆库中取出部分记忆，对待测目标进行训练学习，输出得到RFID标签具体位置，相比传统的室内定位方法，本方法可以连续自动学习并定位，定位精度和定位速度上都有很大的提升。

附图说明

[0040] 图1.本发明一种基于DDPG的RFID室内定位算法总体框架图；

[0041] 图2.本发明一种基于DDPG的RFID室内定位算法流程图

具体实施方式

[0042] 为使本发明的上述目的、特征和优点能够更加明显易懂，下面结合附图和具体实施例对本发明作进一步详细说明。

[0043] 实施例：

[0044] 参见图1，为本发明一种基于DDPG的RFID室内定位算法总体框架图。本发明首先在定位区域对RFID标签进行RSSI数据采集，具体包括：标签反向散射信号，计算机通过数据处理中心发送指令到读写器，读写器进一步控制标签读取，以获取标签的原始RSSI值，并将这些RSSI值输入动作网络和评价网络进行处理。

[0045] 动作网络包括动作估计网络和动作目标网络，动作估计网络利用强化学习中的深度确定性策略逼近行为值函数$Q^\mu(s,a)$和确定性策略$\mu_\theta(s)$，在动作输出方面采用一个网络

来拟合策略函数，直接输出实时动作，实时进行策略梯度更新，可以应对连续动作的输出及大的动作空间，很好的解决了定位连续性的问题，动作目标网络则是用来更新评价网络。评价网络包括评价估计网络和评价目标网络，两者都在输出当前状态的评价，但输入端有所不同：评价估计网络使用最原始的RSSI值施加的动作当做输入，评价目标网络则使用从动作目标网络生成的动作加上状态的观测值加以分析，作为下一状态的目标值。评价网络指导动作网络建立位置记忆库，反向传播，不断梯度更新，最终选出最优RSSI值并输出具体标签位置。

[0046] 参见图2，为本发明一种基于DDPG的RFID室内定位算法流程图。在室内布置若干个读写器和RFID样本标签，具体步骤为：

[0047] 步骤1)初始化：根据RFID定位环境，初始化动作网络$Q(s,a|\theta^{Q})$和评价网络$\mu(s|\theta^{\mu})$，再分别初始化这两个网络的目标网络：$Q'\leftarrow Q$，$\mu'\leftarrow\mu$，初始化记忆库为R；

[0048] 步骤2)执行动作：初始化噪声分布N，构造探索策略μ'，每个动作策略添加一个噪声，执行当前动作a_t并观察当前回报值r_t，然后观察得到下一个状态s_{t+1}，在记忆库R中保存经验(s_t,a_t,r_t,s_{t+1})，并获得当前RFID样本标签的最优RSSI值；

[0049] 步骤3)取记忆并训练：从记忆库R中取出部分随机样本，表示为(s_i,a_i,r_i,s_{i+1})，然后训练更新目标网络，选出最优RSSI值，并将经验存入记忆库，学习过程可表示为

[0050] $y_i=r_i+\gamma Q'(s_{i+1},u'(s_{i+1}|\theta^{u'})|\theta^{Q'})$

[0051] 其中y_i代表目标网络，r_i代表在i学习过程中的回报值，$\theta^{u'}$、$\theta^{Q'}$代表目标权重，γ代表折扣因子；

[0052] 步骤4)根据最小损失函数更新评价网络：

[0053] $$L=\frac{1}{N}\sum_i(y_i-Q(s_i,a_i|\theta^{Q}))^2$$

[0054] 其中L代表损失函数；

[0055] 步骤5)根据策略梯度更新动作网络：

[0056] $$\nabla_{\theta^{\mu}}J\cong\frac{1}{N}\sum_i\nabla_a Q(s,a|\theta^{Q})|_{s=s_i,a=\mu(s_i)}\nabla_{\theta_\mu}\mu(s|\theta^{\mu})|_{s_i}$$

[0057] 其中$\nabla_{\theta^{\mu}}J$代表梯度，用动作的方法调整权重值；

[0058] 步骤6)更新目标网络，即权重更新：

[0059] 评价网络权重更新：$\theta^{Q'}\leftarrow\tau\theta^{Q}+(1-\tau)\theta^{Q'}$；

[0060] 动作网络权重更新：$\theta^{\mu'}\leftarrow\tau\theta^{\mu}+(1-\tau)\theta^{\mu'}$；

[0061] 步骤7)当学习步数结束时，输出样本标签的具体位置，训练完成DDPG定位模型；

[0062] 步骤8)当待测目标进入检测区域时，读写器读取目标所携带的RFID标签的RSSI值，并根据DDPG定位模型配置参数，由于标签在读取过程中，存在信号反射，衰减，多径干扰等因素的影响，应不断学习，并调整学习参数，估算出待测目标的具体坐标值。

[0063] 最后所应说明的是，以上实施例仅用以说明本发明的技术方案而非限制。尽管参照实施例对本发明进行了详细说明，本领域的普通技术人员应当理解，对本发明的技术方案进行修改或者等同替换，都不脱离本发明技术方案的精神和范围，其均应涵盖在本发明的权利要求内。本发明是在多位RFID室内定位技术人员长期进行研究的经验积累基础上，通过创造性劳动而得出，利用动作-评价网络找到最优RSSI值，训练输出样本标签位置，并

将经验存入记忆库，建立DDPG定位网络模型；当有待测目标进入检测区域时，网络模型自动抽取相关记忆，预测并训练得到目标具体位置，有效的解决了室内定位精度低及环境噪声影响等问题，且模型简单，定位成本低。

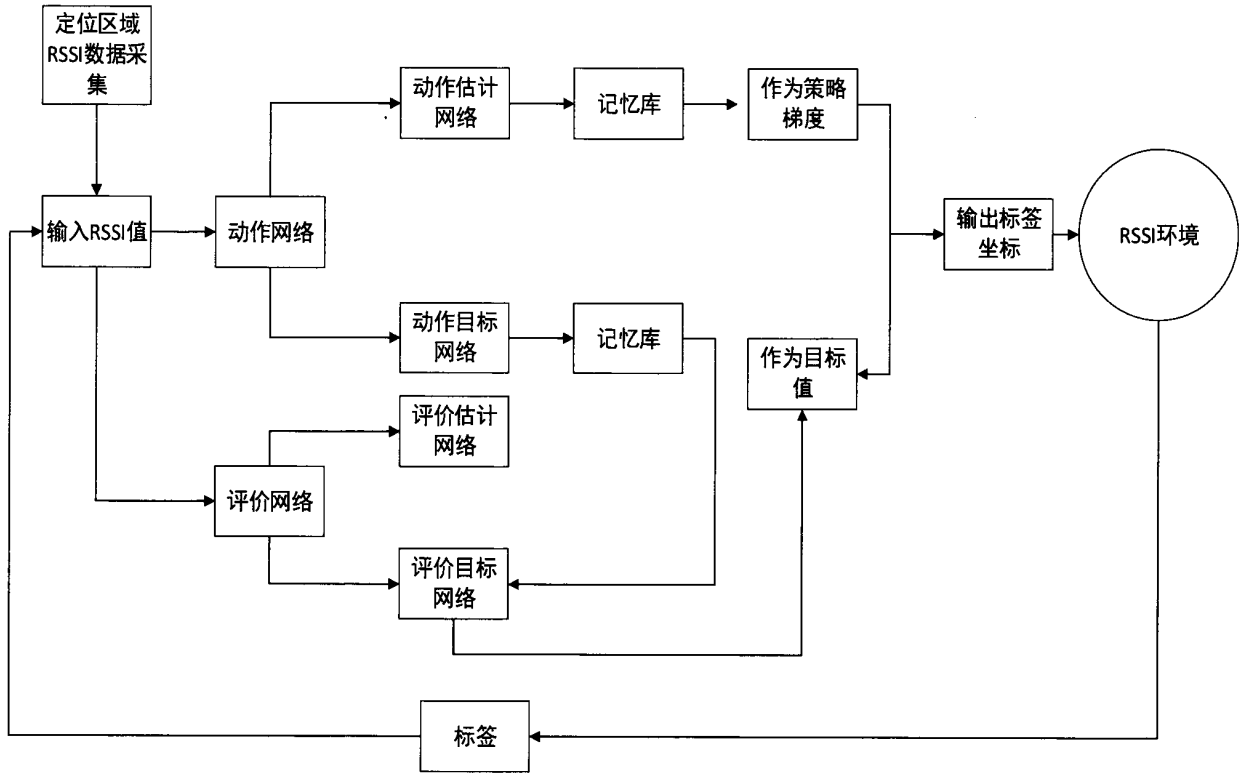

定位区域RSSI数据采集
输入RSSI值
动作网络
动作估计网络
记忆库
作为策略梯度
动作目标网络
记忆库
作为目标值
输出标签坐标
RSSI环境
评价估计网络
评价网络
评价目标网络
标签

图1

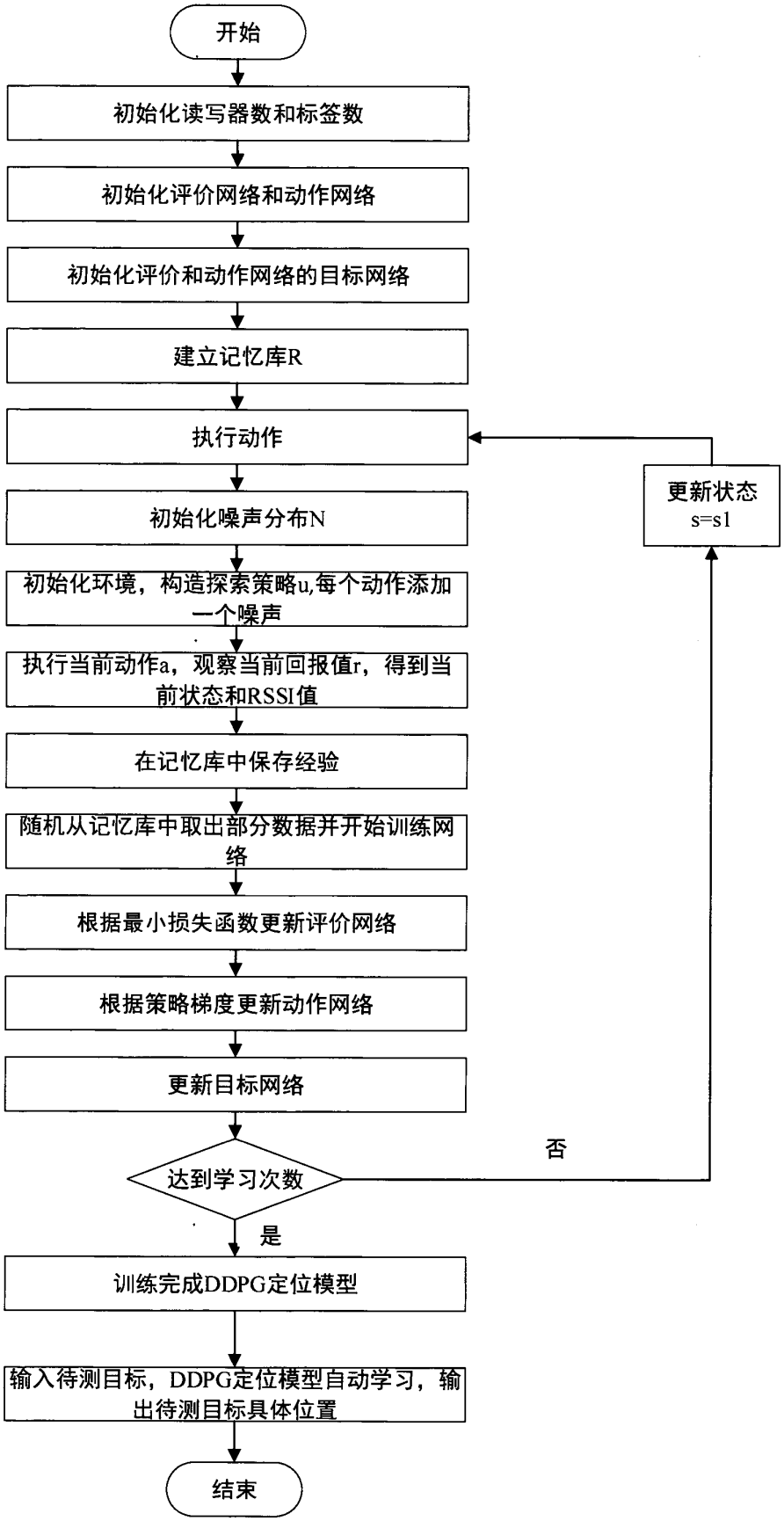

开始
初始化读写器数和标签数
初始化评价网络和动作网络
初始化评价和动作网络的目标网络
建立记忆库R
执行动作
初始化噪声分布N
初始化环境，构造探索策略u,每个动作添加一个噪声
执行当前动作a，观察当前回报值r，得到当前状态和RSSI值
在记忆库中保存经验
随机从记忆库中取出部分数据并开始训练网络
根据最小损失函数更新评价网络
根据策略梯度更新动作网络
更新目标网络
达到学习次数
否
是
更新状态
s=s1
训练完成DDPG定位模型
输入待测目标，DDPG定位模型自动学习，输出待测目标具体位置
结束

图2