计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201100154

能量收集无线通信系统中基于强化学习的能量分配策略

王英恺　王青山
合肥工业大学数学学院　合肥 230001
(2019111237@mail.hfut.edu.cn)

摘　要　随着物联网的普及,对物联网终端设备可使用能量的要求也在提高。能量收集技术拥有广阔前景,其能通过产生可再生能量来解决设备能量短缺问题。考虑到未知环境中可再生能量的不确定性,物联网终端设备需要合理有效的能量分配策略来保证系统持续稳定工作。文中提出了一种基于DQN的深度强化学习能量分配策略,该策略通过DQN算法直接与未知环境交互来逼近目标最优能量分配策略,而不依赖于环境的先验知识。在此基础上,还基于强化学习的特点和系统的非时变系统特征,提出了一种预训练算法来优化该策略的初始化状态和学习速率。在不同的信道数据条件下进行仿真对比实验,结果显示提出的能量分配策略在不同信道条件下均有好于现有策略的性能,且兼具很强的变场景学习能力。

关键词:能量收集;无线通信;能量管理;马尔可夫决策过程;深度强化学习
中图法分类号　TP391

Reinforcement Learning Based Energy Allocation Strategy for Multi-access Wireless Communications with Energy Harvesting

WANG Ying-kai and WANG Qing-shan
School of Mathematics,Hefei University of Technology,Hefei 230001,China

Abstract　Due to the increasing popularization of the Internet of Things (IoT),the requirements for the power that can be used by the terminal equipment of the IoT are also constantly improving. Energy harvesting technology is a promising solution to overcome equipment energy shortages by generating renewable energy. Considering the uncertainty of renewable energy in the unknown environment,the terminal equipment of the IoT needs a reasonable and effective energy allocation strategy to ensure the continuous and stable operation of the system. In this paper,a DQN-based deep reinforcement learning energy allocation strategy is proposed,which uses DQN algorithm to directly interact with the unknown environment to approach the optimal energy allocation strategy without relying on the prior knowledge of the environment. Moreover,a pre-training algorithm is proposed to optimize the initialization state and learning rate of the strategy based on the characteristics of reinforcement learning and time-invariant system. The simulation results under different channel data conditions show that the energy allocation strategy proposed in this paper has better performance than the existing strategy under different channel conditions,and has strong variable scene learning ability.

Keywords　Energy harvesting,Wireless communication,Resource allocation,Markov decision process,Deep reinforcement learning

1　引言

近年来,物联网(IoT)不断普及,其应用范围越来越广泛。但由于物联网终端设备只能携带有限蓄电池,其能量短缺问题始终限制着物联网的进一步发展[1]。能量收集技术(Energy Harvesting,EH)则被认为是一个很有前景的解决方案[2]。EH技术被定义为可以收集环境能量,如太阳能、风能、热能,并将其转换为电能以供设备使用的技术。因此,带EH模块的系统具有一些特别的优势:EH系统在非硬件损坏情况下能够持续收集能量,很大程度地延长了设备使用寿命;在不需要补充能源的前提下,可以被部署在一些平时难以到达的地方[3]。

虽然带有EH模块的系统有以上令人瞩目的优势并得到了广泛的应用[4-5],但目前尚不存在合适的动态能量分配策略来满足EH无线通信系统在未知变化规律或不稳定的可收集能量与信道增益条件下[2,6-9](信道增益描述的是信道自身的传输能力特性)自主工作的要求,因此依旧限制了其进一步的大规模应用。为了解决以上问题,目前已有研究讨论了如何设计基于EH的无线通信系统的最优访问控制策略和能量分配算法[8-12]。具体而言,一种思路[8-12]是通过直接给定EH或信道增益的先验知识分别设计出适用于单用户、多用户EH系统的最优功率分配方法和带有EH模块的MEC系统的最优卸载策略。而另一种思路则是通过假定状态转移函数,使用基于马尔可夫决策过程的动态规划算法[13-15],即通过采用马尔可夫决策过程对系统的功率分配问题建模,进而采用动

到稿日期:2020-11-23　返修日期:2021-02-09
基金项目:国家自然科学基金(61571179)
This work was supported by the National Natural Science Foundation of China(61571179).
通信作者:王青山(qswang@hfut.edu.cn)

态规划方法得到最优的功率分配方案。上述方法均直接或间接地依赖于先验的收集能量的分布和信道增益的分布的系统知识。事实上,这些先验知识在实际使用中是极难获得的,即使通过采样获得一段时间内的环境变化对应的随机分布,后续时间内该随机分布也存在持续变化的可能,使得最终得到的模型无法适应未知环境的情形。

因此,鉴于先验知识问题的难以解决,人们转而寻找一些无模型的基于学习的方法来减小甚至摆脱先验知识的束缚。其中,强化学习便是一种在未知环境中让代理自主学习提高其性能表现而闻名的算法[16]。考虑到强化学习的特性非常契合 EH 系统的特点,本文给出了一些现有在 EH 系统中应用强化学习的例子。文献[17]中,强化学习中的 Q-learning 方法被应用于一个两跳 EH 中继系统以最大化系统吞吐量。文献[18]中,Q-learning 体系则被应用于一个 EH 传感器节点,使得节点在最大化采样率的同时保证能量存储不被耗尽。文献[19]中,EH 无线传感器网络利用 Q-learning 体系实现了在有限先验知识的条件下,达到理想数据包交付率的目标。文献[20-21]采用演员-批评家的深度强化学习方法来设计基于 EH 的无线通信系统的传输方案。文献[22]则考虑了基于 EH 模块的访问控制接入点向多用户传输数据的多访问系统,采用深度强化学习方法设计了能量分配和多访问控制策略。这些研究结果表明,强化学习具有解决未知环境中 EH 系统策略问题的潜质。

本文研究了在未知环境下的 EH 多址无线通信系统的能量分配问题,提出了一种基于 DQN 的强化学习能量分配策略,在没有任何先验系统知识的条件下,直接通过系统与环境的交互来实现在线合理规划和控制 AP 选择多个用户对多个信道的接入,以实现联合协同优化最大化系统长期吞吐量和最大化系统工作时长的系统目标。此外,为了提高系统的变场景学习能力和学习初期的表现,本文基于强化学习和系统的非时变结构特点,提出了一种预学习的改进算法。实验结果表明,本文提出的策略在仿真实验中的表现均优于传统策略,并且预训练算法优化系统的初始状态效果明显,使得策略具有较好的变场景学习能力。

2 系统模型

2.1 信道模型

图 1 所示的系统是一个典型带能量收集的多路无线通信模型,由一个能量收集模块、一个访问控制接入点、一组 u 个用户点和一组 k 个正交频分复用(Orthogonal Frequency Division Multiplexing,OFDM)信道组成。

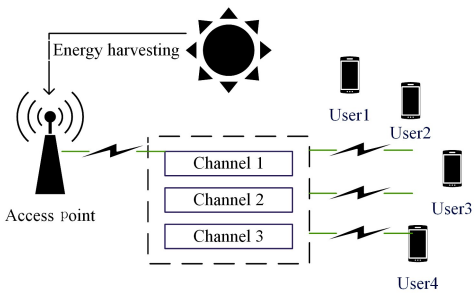


图 1 带能量收集的多路无线通信模型

Fig. 1 Multi-access wireless communications with energy harvesting

在目前的无线通信系统中,实际只支持几个预定义的离散通信模型,并对应于不同的信道编码速率和调制顺序。而不同的传输数据速率,对接收信号功率的要求也不同,通常较高的传输数据率对应于较大的接收信号功率。本文使用 $R=\{r_1,r_2,\cdots,r_M\}$ 表示 M 档传输速率,并用 T_M 表示对应于传输速率 r_M 的最小接收信号功率。

该无线系统是按时隙的形式工作,因此很自然地将第 t 个时隙内的第 k 个用户对应的第 i 个信道的信道增益记为 $\boldsymbol{C}_k^i[t]$(这里的信道增益指传输耗能对应的衰减因子),而 $\boldsymbol{C}[t]$ 则为一个矩阵,维度为 $n\times k$,代表了所有用户点分别对应所有信道数的情况。考虑到时隙的大小通常很小,我们假定在单个时隙内的信道增益是不变的,即每个时隙内对应的信道增益是独立同分布的随机变量,服从于环境模型高度相关的随机分布。而整个时隙序列内的信道增益是时变的。在每个时隙开始时,有多种方法可以获得当前时隙内的信道增益,比较典型的方法是使用脉冲信号,即用户以固定功率向 AP 发送一小段脉冲信号,通过到达的脉冲信号强度估计当前瞬时的信道增益。本文假定脉冲信号使用固定传输功率 P 来传输信号,因此为了支持一个传输速率运行,需要信道增益满足 $P\boldsymbol{C}_k^i[t]>T_M$ 且 $P\boldsymbol{C}_k^i[t]\leqslant T_{M+1}$,从而得到:

$$T_M/P<\boldsymbol{C}_k^i[t]<T_{M+1}/P \tag{1}$$

由式(1)可以直观看出,信道质量 $\boldsymbol{C}_k^i[t]$ 可以完全用最小接收信号功率 T_M 来衡量,进而可以用数据传输速率来衡量信道增益状态。这里需要注意的是,最差的数据传输速率无法被提供,而本文选择的是将无法提供的数据传输速率记为 $r_0=0$。因此,后文均直接使用数据传输速率来描述信道增益。进一步,由数据传输速率服从的分布对应的方差与均值可以用来衡量整体信道状态的质量。

2.2 能量收集模块模型与电池模型

本文假定 EH 模块一直处于工作状态,但同时也考虑了特殊情况下无 EH 模块的工作情形。系统工作的每一个时隙和时隙间的间隔都是相同大小。为了简便起见,认为每一个时隙与该时隙前一个时隙的间隔,即这一段时间内收集的能量都是在该时隙内收集的能量,并记为 $E[t]$。电池有容量上限 $B_{\max}$,每一个时隙开始时的能量记为 $B[t]$。具体地,本文将 $E[t]$ 和 $B[t]$ 均离散化,单位 1 为传输数据包的耗能。这样做是因为通常来说系统的规模是相匹配的,理想状态下每一个时隙收集的能量恰好能支持当前回合的数据传输动作。

与许多其他的工作一样[17],我们将整个能量收集与能量消耗的过程视为一个马尔可夫决策过程。我们将每一个时隙中由于工作而消耗的能量记为 $P[t]$。在每一个时隙的开始阶段,计算上一个时隙收集的能量 $E[t]$,并在该时隙的结束阶段计算该时隙内的工作耗能 $P[t]$,下一个时隙的能量 $E[t+1]$便可以由马尔可夫决策过程计算出:

$$B[t+1]=\min\{B_{\max},E[t]+B[t]-P[t]\} \tag{2}$$

2.3 系统的工作流程

本文假定所有的信号传输都使用二元传输策略,即只有提供完整的服务和不提供服务两种可能,这与文献[2,20]中使用的系统相同。在每个时隙中,一个用户最多只能分配一个信道,每个信道最多只能用于一个用户传递信息。与信道状态矩阵相似,我们将在时隙 t 中的发送动作即通道分配写成与 $\boldsymbol{C}[t]$ 同规模的矩阵 $\boldsymbol{A}[t]$。$\boldsymbol{A}[t]$ 中的元素 $\boldsymbol{A}_k^i[t]$ 满足 $\boldsymbol{A}_k^i[t]\in$

$\{0,1\}$，其中 $\boldsymbol{A}_k^i[t]=0$ 表示不将第 i 个信道分配给第 k 个用户，$\boldsymbol{A}_k^i[t]=1$ 则表示分配。那么自然总发送功率为：

$$P[t]=(\sum_n\sum_k \boldsymbol{A}_k^n[t])P \tag{3}$$

其中，P 是固定发送功率。

如果在时隙开始时，系统内的能量不足以支持该回合选择的发送动作的执行，即 $P[t]>B[t]$，那么系统便会轮空。与绝大多数工作一样，我们也假设激活 AP 的 EH 电路所需的功率与用于信息传输的功率相比可忽略不计，这是因为 EH 电路激活的处理功率通常与实际的传输功率相比非常小。

3 问题公式化和深度 Q-learning 框架

强化学习可以通过特定场景中的自学经验来做出最佳决策，并将所有问题抽象为智能体与环境之间的交互过程来进行建模。在交互过程的每个时间步骤中，智能体接收环境的状态并选择相应的响应动作，然后在下一个时间步骤中，智能体根据环境的反馈来获得奖励值和新的状态以继续学习目标策略，具体流程如图 2 所示。但是，为了使用强化学习，必须先对环境以马尔可夫决策过程的形式建模。

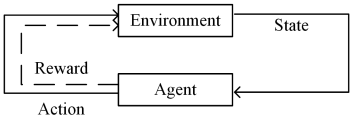


图 2 强化学习流程示意图

Fig. 2 Schematic diagram of reinforcement learning process

3.1 马尔可夫决策过程的形式

马尔可夫决策过程通常可以由一个四元组 (s,a,r,T) 表示，这四元分别为：状态空间 s、动作空间 a、奖励函数 r 与状态转移函数 T。但是，在本文考虑的未知环境条件下，无法给出确定的状态转移函数 T，因此下文只给出状态空间 s、动作空间 a 和奖励函数 r。

(1)状态空间 $s[t]$：由当前时隙的信道状态空间 $(C[t],E[t])$ 组成，即：

$$s[t]\in S=\{C[t],B[t],E[t]\}$$

(2)动作空间 $a[t]$ 与 $C[t]$ 同纬度，用 0，1 表示选择哪一个信道传输。

(3)奖励函数 $r[t]$：

$$r(s[t],a[t])=\begin{cases}\alpha\sum_n\sum_k(a[t]\circ s[t])_n^k, & P[t]\leqslant B[t]\\ -\beta, & P[t]>B[t]\end{cases}$$

而传统的贝尔曼方程迭代方法的局限性也正是在于依赖确定的状态转移函数 T。

3.2 深度强化学习

3.2.1 Q-learning

Q-learning 算法中最核心的概念便是通过动作价值函数 $Q(s,a)$ 来评价对状态 S 执行动作 a 的好坏程度。为了使长期平均奖励最大化，即获得一些最好的动作，Q-learning 算法利用经验 $(s[t],a[t],r[t],s[t+1])$ 学习动作价值函数 $Q(s,a)$。具体来说，就是利用时间差分法中的公式：

$$Q(s,a)\leftarrow Q(s,a)+\alpha[r+\gamma\max_{a'}Q(s',a')-Q(s,a)] \tag{4}$$

对所有经验集合递归地执行上式，直到收敛，其中，$\alpha\in(0,1)$ 为学习率。在所有的 $Q(s,a)$ 产生之后，最优策略——最优动作的集合便自然产生了。而 Q-learning 的一个关键步骤是产生经验 $(s[t],a[t],r[t],s[t+1])$，即学习样本。通常选择的产生方法便是经典的 ε-greedy 算法：

$$a=\begin{cases}\arg\max_a Q(s,a), \text{with probability } \varepsilon\\ \text{randomly chose another action}\end{cases} \tag{5}$$

ε-greedy 算法以一定概率“利用”当前最好的动作，再以一定概率“探索”未选择的动作。传统 Q-learning 使用一张 Q 表以 (S,A) 为行列存储 $Q(s,a)$，但在 $Q(s,a)$ 规模较大时，稀疏的 Q 表既消耗了大量存储空间又造成了探索的低效。因此，随着近年来深度学习的发展，强化学习自然地通过引入神经网络来克服存储大型 Q 表的困难。

3.2.2 深度 Q-learning

深度 Q-learning(Deep Q-network，DQN)的思路是用一组权值去逼近表示 $Q(s,a)$，即 $Q(s,a,\theta)\simeq Q(s,a)$。

通过最小化合适的损失函数 $L_i(\theta_i)$ 来达到用 $Q(s,a,\theta)$ 学习表示 $Q(s,a)$ 的目的。

$$L_i(\theta_i)=\sum_D\mathbb{E}[(R+\gamma\max_{a'}Q(s',a';\theta_{i-1})-Q(s,a;\theta_i))^2] \tag{6}$$

在此基础之上，为 DQN 增加经验池回放和双神经网络(Double DQN，DDQN)错步更新技术，以克服学习样本的高度相关性与分布不平衡导致的神经网络训练困难。经验池回放指将所有的样本放入一个固定大小的经验池中，每次产生新的样本便会更新经验池，每次训练从经验池中取出固定数量批次的样本。双神经网络指使用两个设置完全相同的神经网络，每次由其中一个产生动作，并训练更新另一个网络，间隔一定轮数之后将这个神经网络的参数完全复制到另一个中。因此，DQN 与 DDQN 本质上指代相同，后文不再加以区分。此时的算法框架如图 3 所示。即使使用以上两种技术，神经网络的训练时间依旧较长。

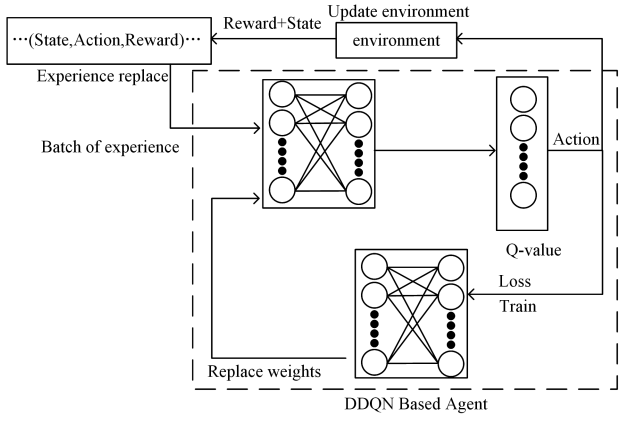


图 3 以 DDQN 为基础的算法框架

Fig. 3 DDQN-based algorithm framework

3.3 预训练模块与总体流程

系统使用变场景时，我们都希望能加速学习的过程以用最短的时间达到使用最优化策略的状态。事实上，考虑的系统有一些特征是独立于环境因素的，如每一个时间的信道增益 $C[t]$、总的系统初始能量 $E[t]$，系统的规模即包含的正交信道数和最大用户数等。

具体来说，动作的选择基于两点：现有能量 $E[t]$ 和当前的信道状态 $C[t]$。如果现有存储在电池中的能量 $E[t]$ 是一

个相对较小的值(几乎是空的),则系统倾向于采用更保守的战略(即尽量选择少或不选择用户进行通信),反之亦然,如果现有存储在电池中的能量 $E[t]$ 几乎是满的,则系统往往采取更为积极的战略。更进一步,若一个用户被选择去传输数据,并不是所有的信道都有待分配的可能性,仅有信道状态最好的那个信道可能会被选择,这不难推广到多个信道组合的情况。因此,一个策略的动作选择应当分为动作选择倾向和具体动作选择这两部分,而这两部分是存在系统可以公共学习的部分,所有可以公共学习的部分以潜在信息的形式被包含于已经学习到的 Q 值中。

基于以上认识,无论采用哪种环境模型,系统都可以从已学习到的 $Q(s,a)$ 部分,通过迁移学习提前学习到一些先验的结构化特征。因此,我们添加一个预训练模块,为 DQN 分配更为合适的预设初始 Q 值 $Q(s,a)$。我们通过让系统在一个更具一般性和代表性的假设环境中学习,得到一组神经网络权重并将其内置在系统中,为我们的目标 DQN 分配合适的初始 Q 值 $Q(s,a)$ 以达到减小对未知动作探索的开销的目的,使得系统能更快地学到一个最优策略。综上所述,本文算法的总体框架如图 4 所示,图 4 中用红框标出预训练模块的部分。

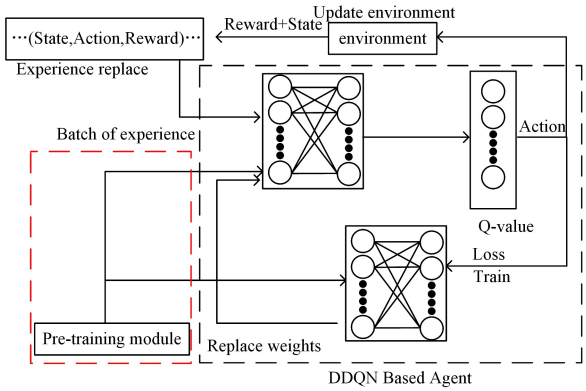


图 4 带有预训练模块的以 DDQN 为基础的算法框架

(电子版为彩色)

Fig. 4 DDQN-based algorithm framework with pre-training module

4 仿真实验

本节将针对提出的算法设置仿真实验,测试其性能并验证其结果。算法的主要评价指标是系统的(平均)长期吞吐量。实验环境所在系统为 Window10,使用的编程语言是 Python3.7,使用的深度学习框架是 Keras。

4.1 仿真实验设置

实验测试的系统参数设置如表 1 所列。

表 1 系统参数设置

Table 1 System parameter settings

Parameter	Value
Number of User	6
Number of Channel	4
Limitation of battery	30\70
Data rate	[0,1,2,3]
EH energy	[1,3,5,7]
The parameters of the reward function	$\alpha=1,\beta=2$

考虑到实验系统对应的动作规模相对来说非常巨大,$|\boldsymbol{A}|=\sum_{i=0}^{4}\boldsymbol{C}_6^i\times\boldsymbol{A}_6^i=1045$,因此实验使用对应较大的神经网络规模,神经网络的具体参数设置如表 2 所列。奖励函数中参数设置为 $\alpha=1,\beta=2$,也就是本文中考虑的奖励函数为:

$$R(S[t],\boldsymbol{A}[t])=\begin{cases}\sum_n\sum_k(\boldsymbol{A}[t]\circ S[t])_n^k, & P[t]\leqslant B[t]\\ -2, & P[t]>B[t]\end{cases}$$

表 2 神经网络参数设置

Table 2 Neural network parameter settings

Parameter	Value
Number of hidden layer	3
Number of hidden layer nodes	256

本文选择使用场景中最典型的 6 个环境模型和 1 个预置模型,其中环境模型 Env1－Env2 对应无 EH 模块的特殊工作情况;Env3－Env6 对应正常工作在不同环境条件与信道条件下的情况;Env7 则为预训练使用的环境模型。通过对应分布进行采样得到具体对应的概率值,如表 3 所列。

表 3 测试环境设置

Table 3 Test environment settings

Environment name	Distribution of channel	Channel probability	Distribution of EH	EH probability	Initial energy
Env1	*Gaussian*($\mu=2,\sigma=1$)	[0.139,0.357,0.360,0.144]	—	—	70
Env2	*Gaussian*($\mu=4,\sigma=1$)	[0.003,0.041,0.280,0.676]	—	—	70
Env3	*Gaussian*($\mu=2,\sigma=1$)	[0.139,0.357,0.360,0.144]	*Poisson*($\lambda=0.85$)	[0.453,0.357,0.143,0.047]	30
Env4	*Gaussian*($\mu=4,\sigma=1$)	[0.003,0.041,0.280,0.676]	*Poisson*($\lambda=0.85$)	[0.453,0.357,0.143,0.047]	30
Env5	*Gaussian*($\mu=2,\sigma=1$)	[0.139,0.357,0.360,0.144]	*Poisson*($\lambda=1.9$)	[0.146,0.291,0.282,0.281]	30
Env6	*Gaussian*($\mu=4,\sigma=1$)	[0.003,0.041,0.280,0.676]	*Poisson*($\lambda=1.9$)	[0.146,0.291,0.282,0.281]	30
Env7	uniform	[0.250,0.250,0.250,0.250]	uniform	[0.250,0.250,0.250,0.250]	30

对比实验则选择一些具有代表性的传统策略:

(1)轮询策略(round-robin)。该策略是最简单的一种负载均衡算法,按照一个已有的固定顺序,以上一次发送的用户为基准,在这一次发送中顺次为后 i 个用户提供服务而不考虑任何的开销与长期累计回报。本文使用 2-轮询策略。

(2)随机策略(random)。该策略只是选择一个随机的动作来执行,即一个动作空间中的随机动作。

(3)最优策略(optimal)。该策略也就是即时最优策略,总是选择在当前时隙内即时回报最大化的行动。换句话说,在状态 s 时,它总是选择以下动作:

$$a=\arg\max_{\forall\boldsymbol{A}\in\mathcal{A}|_{d|_s}}\{\sum_n\sum_k(\boldsymbol{A}\circ S[t])_n^k\}$$

为了正确评价算法的优劣,本文设置如下 3 个性能指标作为评价算法优劣的标准。

(1)正常工作时单位时间内的吞吐量与工作时长,即奖励函数的大小。

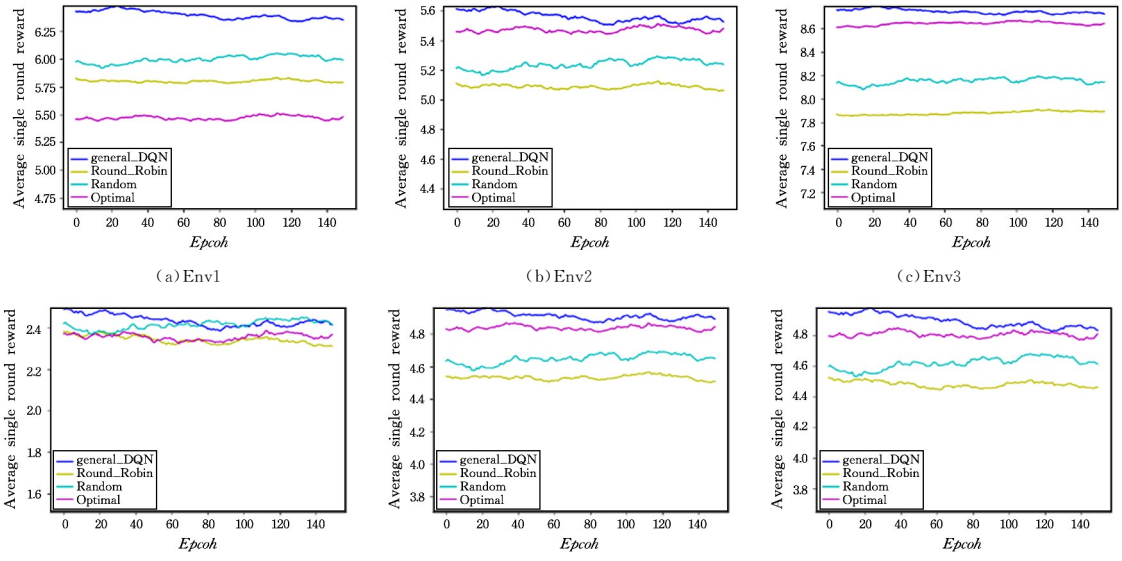
(2)训练初期的表现,即训练初期的奖励函数大小。

(3)变场景学习能力，即在一个新环境中，重新考查以上两个指标。

4.2 仿真实验结果

本节将实验分为 3 部分，第一部分测试策略的性能表现，第二部分测试策略的预学习算法效果，第三部分测试策略的变场景学习能力。

在第一部分实验中，分别在 6 个测试环境模型中将对应已训练好的策略与其他对比策略进行测试，实验结果如图 5(a)—图 5(f)所示。从图中可以看出，本文策略均可以得到最高的奖励函数值。最优策略虽然可以在多数情况下接近本文方法，但是在一些特殊情况(如图 5(e))的环境模型下远低于本文方法得到的奖励函数值。这是由于本文方法的选择是接近于最优选择的，并且可以在能量即将耗尽时选择保守发送，而不至于会轮空系统造成额外的减分。在综合信道状态较差的情况下，所有的算法都有一个比较接近的性能，如图 5(c)所示。综合对比实验结果可知，本文方法可以在系统的长期平均吞吐量与轮空时间的差值这一评价指标下获得一个更好的性能。



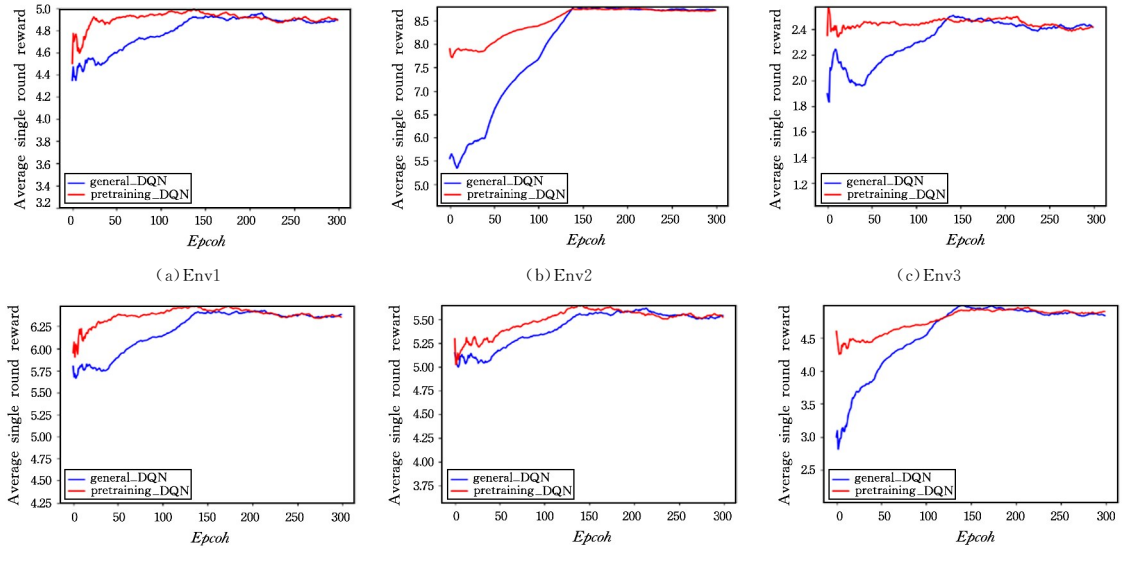
(a)Env1 (b)Env2 (c)Env3

(d)Env4 (e)Env5 (f)Env6

图 5 在各种环境下的性能比较

Fig. 5 Performance comparisons in a variety of environments

在第二部分实验中，分别在 6 个测试环境模型中将使用和不使用预训练算法这两种情况在训练初期奖励值上的区别进行对比，实验结果如图 6 所示。从图中可以看出，所提出的预训练算法可以显著减小训练初期的损失，在训练初期就能达到一个较好的水平，并在一些环境中显著提高了收敛速度。综合对比实验结果可知，预训练算法具有很好的效果。



(a)Env1 (b)Env2 (c)Env3

(d)Env4 (e)Env5 (f)Env6

图 6 在各种环境下的预训练加速效果比较

Fig. 6 Comparison of pre-training acceleration effects in various environments

在第三部分实验中，分别在 6 个测试环境模型中将未训练的策略和使用预训练算法的未训练的策略与传统算法的奖励值进行对比，以考查其变场景的学习能力，实验结果如图 7 所示。

本文将对变场景学习能力的考查分为两步。第一步对模型在不同环境模型中的性能表现进行考查；第二步对模型在不同环境模型中预训练的加速效果进行考查。从图 7 中可以看出，本文策略在环境模型改变之后均能表现出最好或接近最好的性能表现，且具有较好的学习速率提升。综合对比实验结果可知，本文策略对不同环境模型具有较好的变场景学习能力。

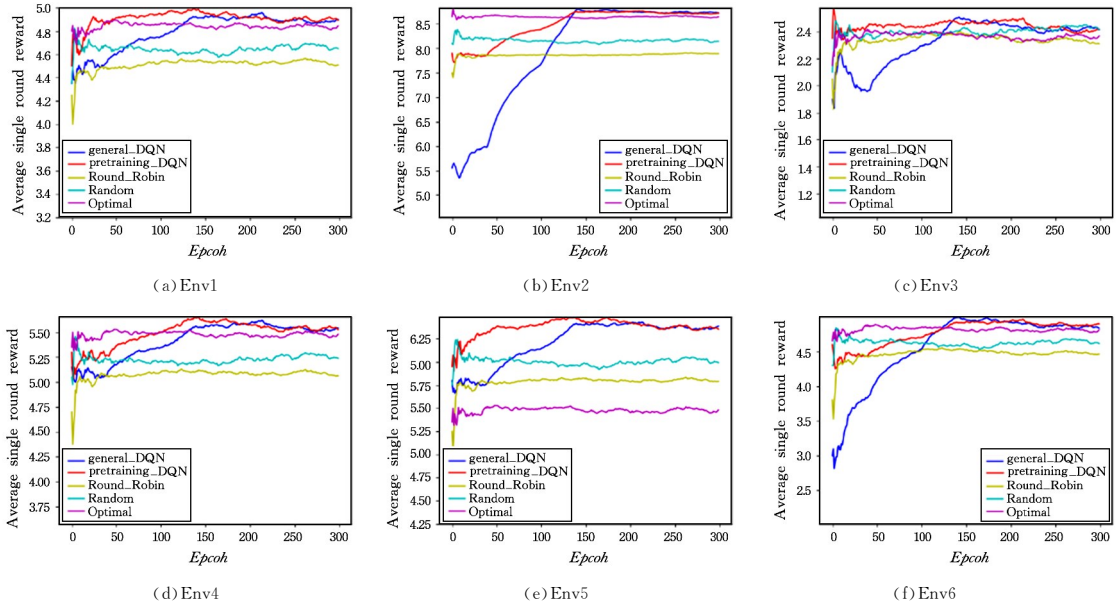


图 7 在各种环境下的变场景能力比较

Fig. 7 Comparison of learning ability of variable scene in various environments

结束语 本文将未知环境中能量收集无线通信系统的能量分配问题看作一个无先验知识的马尔可夫决策过程问题，目标是在最大限度地提高长期平均系统总吞吐量的同时最大限度地增加工作时长。文中使用强化学习中的 DQN 来寻找最优的访问控制策略。更进一步，针对环境模型持续改变的情况，本文基于系统的非时变的结构特性提出了一种有效的预训练算法来加速该策略的收敛速度。实验结果表明，该策略与传统策略相比具有更好的性能，且提出的预训练方法可以显著减小训练初期的损失，使得该策略具有很强的变场景学习能力。

参 考 文 献

[1] YANG Z, XU W, PAN Y, et al. Energy Efficient Resource Allocation in Machine-to-Machine Communications with Multiple Access and Energy Harvesting for IoT[J]. IEEE Internet of Things Journal, 2017, 5(1): 229-245.

[2] ULUKUS S, AYLIN Y, ELZA E, et al. Energy Harvesting Wireless Communications: A Review of Recent Advances[J]. IEEE Journal on Selected Areas in Communications, 2015, 33(3): 360-381.

[3] OZEL O, TUTUNCUOGLU K, ULUKUS S, et al. Fundamental Limits of Energy Harvesting Communications[J]. IEEE Communications Magazine, 2015, 53(4): 126-132.

[4] ZHANG L L, XIONG K, ZHANG Y. UAV-assisted Wireless Energy Harvesting Fog Computing Network Optimization Method[J]. Journal of Software, 2019, 30(1): 9-17.

[5] ABU B, JOSIAH H. Making sense of intermittent energy harvesting[C]// Conference on Embedded Networked Sensor Systems. ACM, 2018: 32-37.

[6] THUC T K, HOSSAIN E, TABASSUM H. Downlink Power Control in Two-Tier Cellular Networks with Energy-Harvesting Small Cells as Stochastic Games[J]. IEEE Transactions on Communications, 2015, 63(12): 5267-5282.

[7] KU M L, LI W, CHEN Y, et al. Advances in Energy Harvesting Communications: Past, Present, and Future Challenges[J]. IEEE Communications Surveys & Tutorials, 2017, 18(2): 1384-1412.

[8] YANG J, ULUKUS S. Optimal Packet Scheduling in an Energy Harvesting Communication System[J]. IEEE Transactions on Communications, 2010, 60(1): 220-230.

[9] TUTUNCUOGLU K, YENER A. Optimum Transmission Policies for Battery Limited Energy Harvesting Nodes[J]. IEEE Transactions on Wireless Communications, 2010, 11(3): 1180-1189.

[10] YUAN F, ZHANG Q T, JIN S, et al. Optimal Harvest-Use-Store Strategy for Energy Harvesting Wireless Systems[J]. IEEE Transactions on Wireless Communications, 2015, 14(2): 698-710.

[11] CHI K K, XU X C, WEI X C. Minimal Base Stations Deployment Scheme Satisfying Node Throughput Requirement in Radio Frequency Energy Harvesting Wireless Sensor Networks[J]. Computer Science, 2018, 45(S1): 332-336.

[12] TIAN X Z, YAO C, ZHAO C, et al. 5G Network oriented Mobile Edge Computation Offloading Strategy[J]. Computer Science, 2020, 47(S2): 286-290.

[13] BLASCO P, GUNDUZ D, DOHLER M. A Learning Theoretic Approach to Energy Harvesting Communication System Optimization[J]. IEEE Transactions on Wireless Communications, 2013, 12(4): 1872-1882.

[14] OZEL O, TUTUNCUOGLU K, YANG J, et al. Transmission with Energy Harvesting Nodes in Fading Wireless Channels: Optimal Policies[J]. IEEE Journal on Selected Areas in Commu-

nications,2011,29(8):1732-1743.

[15] AMIRNAVAEI F,DONG M. Online Power Control Optimization for Wireless Transmission with Energy Harvesting and Storage[J]. IEEE Transactions on Wireless Communications,2016,15(7):4888-4901.

[16] SUTTON R,BARTO A. Reinforcement Learning:An Introduction[M]. MIT Press,1998.

[17] CHU M,LI H,LIAO X,et al. Reinforcement Learning based Multi-Access Control and Battery Prediction with Energy Harvesting in IoT Systems[J]. IEEE Internet of Things Journal,2019,6(2):2009-2020.

[18] FRANCESCO F,BHARATHAN B,RAJESH G. Scaling Configuration of Energy Harvesting Sensors with Reinforcement Learning[C]// Conference on Embedded Networked Sensor Systems. ACM,2018:7-13.

[19] JIA Z G,WANG Z P,HU J T. Work-in-Progress:Q-Learning Based Routing for Transiently Powered Wireless Sensor Network[C]// International Conference on Hardware/Software Codesign and System Synthesis. ACM,2019:1-2.

[20] WEI Y,YU F R,SONG M,et al. User Scheduling and Resource Allocation in HetNets with Hybrid Energy Supply:An Actor-Critic Reinforcement Learning Approach[J]. IEEE Transactions on Wireless Communications,2018,17(1):680-692.

[21] AOUDIA F A,GAUTIER M,BERDER O. RLMan:An Energy Manager based on Reinforcement Learning for Energy Harvesting Wireless Sensor Networks[J]. IEEE Transactions on Green Communications & Networking,2018,2(2):408-417.

[22] CHU M,LI H,LIAO X,et al. Power Control in Energy Harvesting Multiple Access System with Reinforcement Learning[J]. IEEE Internet of Things Journal,2019,6(5):9175-9186.

WANG Ying-kai, born in 1996, postgraduate. His main research interests include reinforcement learning and wireless communication.

WANG Qing-shan, born in 1973, Ph. D supervisor, is a member of China Computer Federation. His main research interests include edge computing and gesture recognition.