基于学习的能量采集认知M2M通信资源分配算法

许艺瀚[1],田永波[1],张扬刚[2],花 敏[1],周 雯[1]
(1. 南京林业大学信息科学技术学院,江苏南京 210037;2. 复旦大学信息科学与工程学院,上海 200433)

摘 要: 本文针对能量采集认知机器到机器(Machine-to-Machine,M2M)通信的能量效率问题,在保证服务质量(Quality of Service,QoS)的条件下,提出了一种能效优化算法.以最大化网络中用户能效为目标,综合考虑传输功率控制、时隙分配、传输模式选择、中继选择以及每个设备的能量状态为约束,将优化问题建模为一个混合整数非线性规划问题.将该能效优化问题转化为离散时间有限状态马尔科夫决策过程(Discrete-time and Finite-state Markov Decision Process,DFMDP)进行求解.提出一种基于深度强化学习的算法寻找最优策略.仿真结果表明,所提算法在平均能效方面优于其他方案,且收敛速度在可接受范围内.

关键词: 能量收集;认知无线电;M2M通信;资源分配;深度强化学习

基金项目: 国家自然科学基金(No.61801225,No.61601275);南京林业大学引进高层次人才和高层次留学回国人员科研基金(No.GXL015)

中图分类号: TN914 **文献标识码:** A **文章编号:** 0372-2112(2023)02-0467-10

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20211258

A Learning-Inspired Resource Allocation for Energy Harvesting-Powered Cognitive M2M Communications

XU Yi-han[1], TIAN Yong-bo[1], ZHANG Yang-gang[2], HUA Min[1], ZHOU Wen[1]
(1. *College of Information Science and Technology, Nanjing Forestry University, Nanjing, Jiangsu* 210037, *China*;
2. *School of Information Science and Technology, Fudan University, Shanghai* 200433, *China*)

Abstract: In order to optimize the energy efficiency for energy harvesting-powered cognitive M2M communications underlaying cellular network, an energy efficient algorithm is proposed while guaranteeing the quality of service of users. Firstly, the problem is formulated as a mixed integer nonlinear programming problem with the goal of maximizing energy efficiency by jointly considering transmission power control, time slot allocation, transmission mode and relay selection with the constraints of the energy status of each device. After that, the optimization problem is modeled as a discrete-time and finite-state Markov decision process. Afterward, a deep reinforcement learning-based algorithm is proposed to find the optimal strategy. Numerical results validate that the proposed scheme outperforms other schemes in terms of average energy efficiency with an acceptable convergence speed.

Key words: energy harvesting; cognitive radio; M2M communications; resource allocation; deep reinforcement learning

Foundation Item(s): National Natural Science Foundation of China (No.61801225, No.61601275); Scientific Research Fund for Introducing High-Level Talents and High-Level Returnees of Nanjing Forestry University (No.GXL015)

1 引言

机器到机器(Machine-to-Machine,M2M)通信作为一种前沿的物联网技术,引起了工业界和学术界的广泛关注.与传统的人与人(Human-to-Human,H2H)通信不同,M2M通信有望在不需要人为干预的情况下,实现各种异构设备间的随时随地连接[1,2].然而,M2M通信涉及大量并发接入需求,进而加剧了频谱稀缺和高能耗问题的严重性.虽然第三代合作伙伴计划(Third Generation Partnership Project,3GPP)持续推进下一代通信技术以缓解日益严重的频谱不足与能耗问题,但对于在异构环境下为大量设备提供不同服务质量(Quality of Service,QoS)需求的业务,其资源分配策略

收稿日期:2021-09-13;修回日期:2022-08-28;责任编辑:孙瑶

尚未得到很好的解决. 与此同时,文献[3~7]针对M2M通信中资源分配问题进行了研究,但这些前期工作主要集中在网络性能方面,如丢包率、时延和吞吐量,很少考虑对能耗和H2H通信的影响[8~10]. 因此,为M2M通信设计一种高能效且具有强抗干扰能力的资源分配策略显得尤为重要.

认知M2M通信是一种将认知无线电(Cognitive Radio,CR)集成于M2M通信中的新技术,认知M2M通信使设备能够从环境中学习并利用未被占用的授权频谱来提高频谱效率,同时避免对H2H通信的干扰. 除了提高频谱效率,M2M通信还能提高能量效率. M2M通信作为物联网技术的关键,涉及了大量的传感器类设备,这些设备具有电池能量受限且难以频繁更换等缺点. 因此,能量采集(Energy Harvesting,EH)技术将成为一种极具吸引力的解决方案. 然而,由于环境能量的波动性以及能量转换技术的不成熟,每个设备的可利用能量将成为能量采集认知M2M网络(Energy Harvesting-powered Cognitive M2M Networks,EH-CMNs)中资源分配策略设计的重要因素之一.

为了解决这个问题,本文为EH-CMNs提出了一种高能效的资源分配策略. 该策略的目的在于综合考虑传输功率控制、时隙分配、传输模式选择、中继选择以及每个设备的能量状态来最大化EH-CMNs的平均能效. 该优化问题为一个非凸混合整数非线性规划问题,传统的凸优化算法无法直接解决该问题. 为此,本文将原始的优化问题建模为一个离散时间有限状态马尔可夫决策过程(Discrete-time and Finite-state Markov Decision Process,DFMDP),其中每个设备被假设为一个agent,要求在事先无须完整的网络全局信息,仅有本地信息的情况下,能够有效地与环境互动,从而自适性地学习到最优分配策略. 因此,提出了一种深度强化学习(Deep Reinforcement Learning,DRL)算法来寻求最优分配策略. 实验表明,所提算法在能效方面优于其他算法,且收敛速度在可接受范围.

2 系统模型描述与问题建模

2.1 网络模型

本文考虑了一个EH-CMN场景,它由一个位于小区中心、覆盖半径为R的基站(Base Station,BS),N个蜂窝用户(Cellular Users,CUs)记为$c_i(i\in\{1,2,\cdots,N\})$和$M$个M2M通信设备对记为$d_j(j\in 1,2,\cdots,M\})$组成. 蜂窝用户和M2M通信设备随机分布在覆盖区域. 每个M2M对都有一个发送端(DU_Tx)和接收端(DU_Rx). 简单起见,本文仅考虑M2M设备配备了EH功能,CUs仍由传统电池供电. 图1为网络模型场景图. 此外,为了提高资源利用率,本文假设设备支持两种通信模式:协作传输和直接传输. 在协作传输模式中,令该模式仅支持两跳传输,中继设备用DU_Rly表示. 本文首先定义了一个二元变量$\alpha_{d_j}\in\{0,1\},j\in(1,2,\cdots,M)$来指示第$j$个设备使用的传输模式,$\alpha_{d_j}=1$表示第$j$个设备处于直接传输模式,$\alpha_{d_j}=0$则表示第$j$个设备处于协作传输模式. 媒体接入层(Media Access Control,MAC)采用(Time Division Multiple Access,TDMA)接入模式,每个传输帧可以被划分为多个时隙,以供设备使用. 假设每个传输帧包括K个时隙,时隙集表示为$\psi=(1,2,\cdots,K)$, $t_0=0$, $t_K=T$,且每个时隙的持续时间为$\tau_k=t_k-t_{k-1}$, $\forall k\in\psi$.

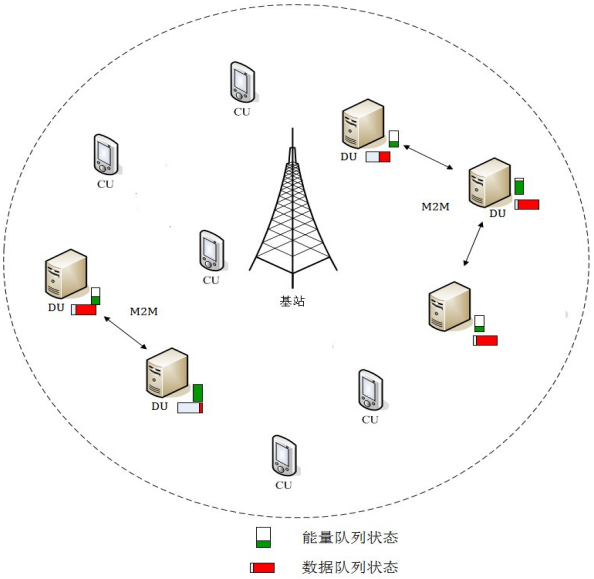


图1 网络场景图

对于直接传输,本文定义另一个二元参数$\beta_{d_j}^k\in\{0,1\}$, $\left(j\in(1,2,\cdots,M),\ \forall k\in\psi\right)$来表示分配给某一设备的时隙. 其中$\beta_{d_j}^k=1$表示第$j$个设备分配在第$k$个时隙进行直接传输,$\beta_{d_j}^k=0$则表示第$k$个时隙没有被分配给第$j$个设备进行直接传输. 该模型还做了另外两个合理的假设:(1)每个设备在一个时隙内只能从另外一个设备接收数据;(2)在一个时间帧内,每个设备最多只能被分配一个时隙进行传输. 这两个假设目的在于为了保证每一设备传输机会的公平性. 因此,可以得到两个约束条件:

$$\sum_{j=1}^{M}\beta_{d_j}^k\leqslant 1,\ k\in\psi \tag{1}$$

$$\sum_{k=1}^{K}\beta_{d_j}^k\leqslant 1,\ j\in(1,2,\cdots,M) \tag{2}$$

对于协作传输,本文假设传输帧中共有K个时隙可以分配给DU_Tx-DU_Rly链路的同时,也可以分配给DU_Rly-DU_Rx链路. 该假设主要用于保证直接传输和

协作传输之间的公平性,从而获得最优的资源分配策略. 同样,本文额外定义了一个二元标识符 $\delta^k_{d_j\to d_r}\in\{0,1\},\ (j,r\in(1,2,\cdots,M),\forall k\in\psi)$ 来指示是否将第k个时隙分配给第j个设备,用于向第r个设备传输数据. 其中,第r个设备为第j个设备的中继. $\delta^k_{d_j\to d_r\to d_z}\in\{0,1\},\ (j,r,z\in(1,2,\cdots,M),\forall k\in\psi)$ 用来表示第r个设备将在第k个时隙,转发来自第j个设备的数据到第z个设备. 该模型假设每个DU_Tx在传输帧的任何时隙中只能选择一个DU_Rly,且每个DU_Rly在传输帧的任何时隙中只能转发来自一个DU_Tx的数据. 因此,可以得到两个约束条件:

$$\sum_{r=1,r\neq j}^{M}\delta^k_{d_j\to d_r}\leqslant 1,\quad \sum_{j=1,j\neq r}^{M}\delta^k_{d_j\to d_r}\leqslant 1 \tag{3}$$

$$\sum_{j=1,j\neq r}^{M}\delta^k_{d_j\to d_r\to d_z}\leqslant 1,\quad \sum_{r=1,r\neq j}^{M}\delta^k_{d_j\to d_r\to d_z}\leqslant 1 \tag{4}$$

此外,由于每条链路最多只能分配一个时隙,所以式(5)应得到满足,即

$$\sum_{k=1}^{K}\delta^k_{d_j\to d_r}\leqslant 1,\quad \sum_{k=1}^{K}\delta^k_{d_j\to d_r\to d_z}\leqslant 1,\ j\neq r \tag{5}$$

另一个需要注意的方面是,从DU_Tx到DU_Rly的数据传输应该先于从DU_Rly到DU_Rx的数据传输,所以可得约束条件:

$$\sum_{k=1}^{x}\delta^k_{d_j\to d_r}-\sum_{k=x+1}^{K}\delta^k_{d_j\to d_r\to d_z}\geqslant 0,\ x\in(1,2,\cdots,K-1) \tag{6}$$

2.2　数据传输模型

在本模型中,蜂窝网络中的每个CU都预先分配了上行频谱资源,其带宽为B,且相互正交. 假设每个认知M2M对可以在时间上作为次要用户复用已分配给CU的上行链路频谱,则可以推导出第i个CU的瞬时信号与干扰加噪声比(Signal-to-Interference plus Noise Ratio,SINR):

$$\mathrm{SINR}_{c_i,k}=\frac{p_{i,k}\cdot g^k_{c_i-\mathrm{BS}}}{\sum\limits_{d_j\in M}p_{j,k}\cdot g^k_{d_j-\mathrm{BS}}+n_0} \tag{7}$$

根据香农公式,可得第i个CU的瞬时传输率为

$$R_{c_i,k}=B\cdot\log_2(1+\mathrm{SINR}_{c_i,k}) \tag{8}$$

因此可以得到CUs的长期平均传输率为

$$R_c=\lim_{K\to\infty}\sup\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N}E[R_{c_i,k}] \tag{9}$$

其中,$p_{i,k}$和$p_{j,k}$分别为第i个CU和第j个M2M设备在第k个时隙的瞬时发射功率;g^k表示i,j和BS之间的信道增益;n_0为噪声功率,等于$B\cdot\rho_n$,其中ρ_n是噪声密度函数. 此外,为了保证CUs的传输速率,R_c的值应达到最小传输速率阈值$\mathrm{TR}_{\mathrm{th}}$.

在M2M通信中,不同功能的设备可能有不同的传输速率要求. 在本文的网络模型中,将R_j表示为第j个设备的传输速率,它可表示为

$$R_j=\alpha_{d_j}\cdot R^d_j+(1-\alpha_{d_j})\cdot R^c_j,\ j\in(1,2,\cdots,M) \tag{10}$$

其中,R^d_j为直接传输模式下第j个设备的传输速率;R^c_j为通过中继设备向目的终端传输数据时第j个设备的传输速率. 根据上述分析,可以推导出直接传输和协作传输的瞬时SINR. 式(11)、式(12)和式(13)分别为第k个时隙中直接链路、DU_Tx-DU_Rly链路和DU_Rly-DU_Rx链路的瞬时SINR.

$$\mathrm{SINR}^d_{j,k}=\frac{p^d_{j,k}\cdot g_{d_j-d_z}}{\sum\limits_{j_1=1,j_1^1jr=1,r^1j,j_1}^{M}\sum\limits_{r=1,r^1j,j_1}^{M}\delta^k_{d_{j_1}\to d_r}\cdot p^s_{j_1,r,k}\cdot g_{d_{j_1}-d_r}+n_0} \tag{11}$$

其中,$p^d_{j,k}$表示第j个设备在第k个时隙向作为目的终端的设备z传输数据时的瞬时发射功率;$g_{d_j-d_z}$是设备j和设备z之间的信道增益;$p^s_{j_1,r,k}$表示第k个时隙中第j_1个设备向第r个设备传输数据的瞬时发射功率,设备r被选为设备j_1的中继;$g_{d_{j_1}-d_r}$是设备j_1和设备r之间的信道增益.

$$\begin{aligned}\mathrm{SINR}^{s\to r}_{j,r,k}&=\frac{p^{s\to r}_{j,r,k}\cdot g_{d_j-d_r}}{I^{s\to r}_{j,r,k}+n_0}\\ I^{s\to r}_{j,r,k}&=\sum_{\substack{j_1=1\\ j_1\neq j,r}}^{M}\sum_{\substack{r_1=1\\ r_1\neq j,j_1,r}}^{M}\delta^k_{d_{j_1}\to d_{r_1}}\cdot p^{s\to r}_{j_1,r_1,k}\cdot g_{d_{j_1}-d_r}\\ &\quad+\sum_{\substack{j_1=1\\ j_1\neq j,r}}^{M}\beta^k_{d_{j_1}}\cdot p^d_{j_1,k}\cdot g_{d_{j_1}-d_r}\\ &\quad+\sum_{\substack{j_1=1\\ j_1\neq j,r}}^{M}\sum_{\substack{r_1=1\\ r_1\neq j,j_1,r}}^{M}\delta^k_{d_{j_1}\to d_{r_1}\to d_z}\cdot p^{r\to d}_{j_1,r_1,k}\cdot g_{d_{r_1}\to d_z}\end{aligned} \tag{12}$$

其中,$p^{s\to r}_{j,r,k}$表示第j个设备向第r个设备发送数据时的瞬时发送功率,设备r在第k个时隙被选为中继;$p^{r\to d}_{j,r,k}$表示第j个设备在第k个时隙时向目的终端发送数据时,设备r的瞬时发送功率;$I^{s\to r}_{j,r,k}$的表达式分为三项,第一项表示来自其他DU_Tx-DU_Rly链路的干扰,第二项表示来自DU_Tx和DU_Rx之间的传输干扰,第三项表示来自DU_Rly-DU_Rx链路的干扰.

$$\begin{aligned}\mathrm{SINR}^{r\to d}_{j,r,k}&=\frac{p^{r\to d}_{j,r,k}\cdot g_{d_r-d_z}}{I^{r\to d}_{j,r,k}+n_0}\\ I^{r\to d}_{j,r,k}&=\sum_{\substack{j_1=1\\ j_1\neq r}}^{M}\sum_{\substack{r_1=1\\ r_1\neq r,j_1}}^{M}\delta^k_{d_{j_1}\to d_{r_1}}\cdot p^{s\to r}_{j_1,r_1,k}\cdot g_{d_{j_1}-d_r}\end{aligned} \tag{13}$$

其中,$I^{r\to d}_{j,r,k}$为在第k个时隙选择设备r作为设备j的中继时,中继设备与目的设备之间的总瞬时干扰. 根据香农公式,可得直接链路的传输速率为

$$R_j^d=\sum_{k=1}^{K}\beta_{d_j}^k\cdot B\cdot\log_2\left(1+\mathrm{SINR}_{j,k}^d\right)\tag{14}$$

协作模式的传输速率R_j^c可以分为两个部分:一部分是DU_Tx-DU_Rly链路的传输速率$R_j^{c,s\to r}$;另一部分是DU_Rly-DU_Rx链路的传输速率$R_j^{c,r\to d}$:

$$R_j^{c,s\to r}=\sum_{\substack{r=1\\r\neq j}}^{M}\sum_{k=1}^{K}\beta_{d_j\to d_r}^k\cdot B\cdot\log_2\left(1+\mathrm{SINR}_{j,r,k}^{s\to r}\right)\tag{15}$$

$$R_j^{c,r\to d}=\sum_{\substack{r=1\\r\neq j}}^{M}\sum_{k=1}^{K}\beta_{d_j\to d_r\to d_z}^k\cdot B\cdot\log_2\left(1+\mathrm{SINR}_{j,r,k}^{r\to d}\right)\tag{16}$$

然而,在协同传输下,DU_Tx和DU_Rx的传输速率受到DU_Tx-DU_Rly链路和DU_Rly-DU_Rx链路之间较小传输速率的限制.因此,协作模式的传输速率应为$R_j^c=\min\left(R_j^{c,s\to r},R_j^{c,r\to d}\right)$.

2.3 数据服务模型

在本场景下,假设数据以分组的形式存储在设备缓冲区中.每个设备上的数据遵循独立同分布(independently and identically distributed,i. i. d.)序列,平均速率为λ_d[11].实际上,本模型假设设备的缓冲区是有限的且以先进先出的方式提供服务,将$\mathrm{DQ}_{d_j}^k$表示为时隙k中设备j的瞬时数据队列长度.设备的最大数据队列长度用$\mathrm{DQ}_{d_j}^{\max}$表示,则可得瞬时数据队列长度的更新为

$$\mathrm{DQ}_{d_j}^k=\min\left\{DQ_{d_j}^{\max},DQ_{d_j}^{k-1}\right.$$
$$\left.-\min\left\{\left\lfloor\frac{\alpha_{d_j}\cdot R_j^d+\left(1-\alpha_{d_j}\right)\cdot R_j^c}{\mathrm{PS}_{\mathrm{data}}}\cdot\tau_k\right\rfloor,DQ_{d_j}^{k-1}\right\}+A_{d_j}^{k-1}\right\}\tag{17}$$

其中,$\mathrm{PS}_{\mathrm{data}}$为分组大小,单位为比特/分组;$\left\lfloor\alpha_{d_j}\cdot R_j^d+\left(1-\alpha_{d_j}\right)\cdot R_j^c\right\rfloor\Big/\left(\mathrm{PS}_{\mathrm{data}}\right)\cdot\tau_k$为时隙$k-1$中设备$j$的传输链路的瞬时服务分组数;$A_{d_j}^{k-1}$为时隙$k-1$内设备$j$的到达分组数.

2.4 能量收集模型

在本模型中,本文并没有直接考虑能量收集、转化效率以及能效是否随设备的移动性而发生变化等问题.类似地,本文假设能量也是以能量分组的形式收集并存储于设备的电池中,$E_{j,k}$表示设备j在第k个时隙中收集到的能量,$\left\{E_{j,1},E_{j,2},\cdots,E_{j,t},\cdots,E_{j,K}\right\}$可表示在传输帧中收集的能量的时间序列,它是平均速率为λ_e的i. i. d序列.$\mathrm{EQ}_{d_j}^k$表示第k个时隙中设备j的瞬时能量队列长度,设备的最大能量队列长度用$\mathrm{EQ}_{d_j}^{\max}$表示.因此,可得瞬时能量队列长度的更新为

$$\mathrm{EQ}_{d_j}^k=\min\left\{\mathrm{EQ}_{d_j}^{\max},\mathrm{EQ}_{d_j}^{k-1}-\right.$$
$$\left.\min\left\{\left\lfloor\frac{p_{j,k-1}}{\mathrm{PS}_{\mathrm{energy}}}\cdot\tau_k\right\rfloor,\mathrm{EQ}_{d_j}^{k-1}\right\}+E_{j,k-1}\right\}\tag{18}$$

其中,$\mathrm{PS}_{\mathrm{energy}}$表示能量包大小,单位为焦耳/分组;$p_{j,k-1}$表示设备在$k-1$个时隙的发射功率.根据传输模式,$p_{j,k-1}$可以设置为$p_{j,k-1}^d$,$p_{j,k-1}^{s\to r}$,$p_{j,k-1}^{r\to d}$.

需要注意的是,由于电池容量是有限的,由式(18)可推导出两个约束条件,即

$$\sum_{k=1}^{K}\left\lfloor\frac{p_{j,k-1}}{\mathrm{PS}_{\mathrm{energy}}}\cdot\tau_k\right\rfloor\leqslant\sum_{k=1}^{K}\mathrm{EQ}_{d_j}^k,\quad\forall K\in\{1,2,\cdots\}\tag{19}$$

$$\sum_{k=1}^{K}\mathrm{EQ}_{d_j}^k-\sum_{k=1}^{K}\left\lfloor\frac{p_{j,k-1}}{\mathrm{PS}_{\mathrm{energy}}}\cdot\tau_k\right\rfloor\leqslant\mathrm{EQ}_{d_j}^{\max},\quad\forall K\in\{1,2,\cdots\}\tag{20}$$

其中,式(19)表示当前可用能量不能超过电池的总能量;式(20)表示电池中存储的能量不能超过最大电池容量.

2.5 能效模型

在本模型中,本文将EH-CMNs的能效定义为传输速率与消耗的传输功率之比.式(21)为时隙k中第j个设备的能效:

$$\mathrm{EE}_{d_j}^k=\frac{\alpha_{d_j}\cdot R_j^d+(1-\alpha_{d_j})\cdot R_j^c}{p_{j,k}},\quad\forall j\in\{1,2,\cdots,M\},\forall j\in\psi\tag{21}$$

因此,整体EH-CMNs的平均能效为

$$\mathrm{EE}=\frac{1}{M}\cdot\sum_{k=1}^{K}\sum_{j=1}^{M}\mathrm{EE}_{d_j}^k\tag{22}$$

相应的EE最大化问题可表示为

$$\underset{\alpha_{d_j},\beta_{d_j}^k,\delta_{d_j}^k,p_{j,k}}{\text{maximize}}\ \mathrm{EE}\tag{23}$$

$$\sum_{j=1}^{M}\beta_{d_j}^k\leqslant1,\quad\sum_{k=1}^{K}\beta_{d_j}^k\leqslant1,\quad k\in\psi,j\in(1,2,\cdots,M)\tag{23a}$$

$$\sum_{r=1,r\neq j}^{M}\delta_{d_j\to d_r}^k\leqslant1,\quad\sum_{j=1,j\neq r}^{M}\delta_{d_j\to d_r}^k\leqslant1\tag{23b}$$

$$\sum_{j=1,j\neq r}^{M}\delta_{d_j\to d_r\to d_z}^k\leqslant1,\quad\sum_{r=1,r\neq j}^{M}\delta_{d_j\to d_r\to d_z}^k\leqslant1\tag{23c}$$

$$\sum_{k=1}^{K}\delta_{d_j\to d_r}^{k}\leqslant 1,\quad \sum_{k=1}^{K}\delta_{d_r\to d_z}^{k}\leqslant 1,\ j\neq r \tag{23d}$$

$$\sum_{k=1}^{x}\delta_{d_j\to d_r}^{k}-\sum_{k=x+1}^{K}\delta_{d_j\to d_r\to d_z}^{k}\geqslant 0,\quad x\in(1,2,\cdots,K-1) \tag{23e}$$

$$\lim_{K\to\infty}\sup\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N}E\left[R_{c_i,k}\right]\geqslant \mathrm{TR}_{\mathrm{th}} \tag{23f}$$

$$\sum_{k=1}^{K}\left[\frac{p_{j,k-1}}{PS_{\mathrm{energy}}}\cdot\tau_k\right]\leqslant\sum_{k=1}^{K}\mathrm{EQ}_{d_j}^{k},\quad \forall K\in\{1,2,\cdots\} \tag{23g}$$

$$\sum_{k=1}^{K}\mathrm{EQ}_{d_j}^{k}-\sum_{k=1}^{K}\left[\frac{p_{n,k-1}}{\mathrm{PS}_{\mathrm{energy}}}\cdot\tau_k\right]\leqslant\mathrm{EQ}_{d_j}^{\max},\quad \forall K\in\{1,2,\cdots\} \tag{23h}$$

$$p_{j,k}^{d}\leqslant p_j^{\max}\ \forall j\in(1,2,\cdots,M),\ \forall k\in\psi \tag{23i}$$

$$p_{j,r,k}^{s\to r}\leqslant p_j^{\max}\ j,r\in(1,2,\cdots,M),j\neq r,\ \forall k\in\psi \tag{23j}$$

$$p_{j,r,k}^{r\to d}\leqslant p_j^{\max}\ j,r\in(1,2,\cdots,M),j\neq r,\ \forall k\in\psi \tag{23k}$$

其中,式(23a)表示每个设备在一个时隙内只能从另外一个设备接收数据,且在一个时间帧内,每个设备最多只能被分配一个时隙进行传输;式(23b)和式(23c)用于表示每个DU_Tx在传输帧的任何时隙内只能选择一个DU_Rly,且每个DU_Rly在任何时隙内只能转发来自一个DU_Tx的数据;式(23d)表示每条链路同时最多只能被分配一个时隙;式(23e)表示DU_Tx到DU_Rly的数据传输应该先于从DU_Rly到DU_Rx的数据传输;式(23f)用于保证CU的数据传输速率不小于其要求的最小阈值速率;式(23g)和式(23h)分别用于限制设备的当前可用能量不能超过电池的总能量以及电池中存储的能量不能超过最大电池容量;式(23i)、式(23j)和式(23k)分别表示在直接传输模式和协作传输模式下各发射机的发射功率不能大于设备的最大发射功率.

3　DFMDP模型与优化算法

从能效最大化问题可以看出,这是一个多目标优化问题. 同时,由于变量$p_{n,k}$是连续的,而$\alpha_{d_j},\beta_{d_j}^{k},\delta_{d_j}^{k}$是二元的,所以式(23)是一个混合整数非线性规划问题,传统的凸优化方法无法直接求解. 即使将原问题转化为一个可处理的凸优化问题,该问题仍然需要先验网络信息. 此外,从式(17)和式(18)可以发现数据和能量分组都只与当前到达和先前剩余有关. 因此,可以将式(23)表述为一个DFMDP[12].

3.1　DFMDP模型

通常,DFMDP由四元素(S,A,p,r)组成,其中S是有限状态集,A是有限动作集,p是从状态s到状态s'的转移概率$(\forall s\in S,\forall s'\in S)$,在动作$a$$(\forall a\in A)$被执行后,$r$是$a$$(\forall a\in A)$被执行后获得的即时奖励. 本文假设了M2M设备为agent,用π表示从状态到动作的映射策略,目标是为了找到最优策略π^{*},以在DFMDP的有限时间内最大化奖励函数. 因此,本文所提出的模型的详细四元素设计如下.

状态　状态是一组特定的离散或连续变量,agent可以从与环境的交互中感知这些变量. 在本文的场景中,第j个DU在第k个时隙的状态可以用$s_{d_j}^{k}\in S$表示,$s_{d_j}^{k}$包含两个部分:

(1) 在第k个时隙开始时的第j个DU的数据流队列长度$\mathrm{DQ}_{d_j}^{k}$,记为$s_{d_j}^{k}(1)$;

(2) 在第k个时隙开始时的第j个DU的能量队列长度$\mathrm{EQ}_{d_j}^{k}$,记为$s_{d_j}^{k}(2)$.

为保证状态空间探索的完整性,将$\mathrm{DQ}_{d_j}^{k}$和$\mathrm{EQ}_{d_j}^{k}$指定为整数,分别取$(0,1,\cdots,\mathrm{DQ}_{d_j}^{\max})$和$(0,1,\cdots,\mathrm{EQ}_{d_j}^{\max})$. 因此,在特定时隙内的网络状态可以表示为$s_{d_j}^{k}=\left\{s_{d_j}^{k}(1),s_{d_j}^{k}(2)\right\}$. 此外,整个网络状态空间可由式(24)表示.

$$\begin{aligned}&\boldsymbol{S}=[s_1,s_2,\cdots,s_k,\ldots,s_K]^{\mathrm{T}}\\&\boldsymbol{s}_k=\left(s_1^{k}(1),s_1^{k}(2),s_2^{k}(1),s_2^{k}(2),\cdots,s_M^{k}(1),s_M^{k}(2)\right)\end{aligned} \tag{24}$$

动作　动作是agent可以执行的一组动作来响应不同的状态. 同样,本文用$a_{m,k}$表示在第k个时隙内,第j个DU所采取的动作. $a_{m,k}$在本模型中由4个部分组成:

(1) 在第k个时隙开始时的第j个DU的传输模式$\alpha_{d_j}(k)$;

(2) 在第k个时隙开始时的第j个DU的时隙分配情况$\beta_{d_j}(k)$;

(3) 在第k个时隙开始时的第j个DU的中继选择情况$\delta_{d_j}(k)$;

(4) 在第k个时隙开始时的第j个DU的传输功率$p_j(k)$.

为确保动作空间探索的完整性,$p_{j,k}^{d},p_{j,r,k}^{s\to r},p_{j,r,k}^{r\to d}$应受限于最大发射功率$p_j^{\max}$. 因此,在第$k$个时隙内的第$j$个DU应采取的动作为$a_{m,k}=\left\{a_{m,k}^{(1)},a_{m,k}^{(2)},a_{m,k}^{(3)},a_{m,k}^{(4)}\right\}$. 由此可得,DU的整个动作空间如式(25)所示:

$$\begin{aligned}&\boldsymbol{A}=[a_1,a_2,\cdots,a_k,\cdots,a_K]^{\mathrm{T}}\\&\boldsymbol{a}_k=\left(a_{1,k}^{(1)},a_{1,k}^{(2)},a_{1,k}^{(3)},a_{1,k}^{(4)},\cdots,a_{M,k}^{(1)},a_{M,k}^{(2)},a_{M,k}^{(3)},a_{M,k}^{(4)}\right)\end{aligned} \tag{25}$$

$r(k)$奖励　$r(k)$表示agent在执行一个动作a_k后的即时奖励. 根据最大化EH-CMNs的平均能效的目标,将奖励定义为相应的优化问题,如式(23)所示. 因此,由时间步长k可得累计折扣奖励$R(k)$为

$$R(k)=\sum_{K=k}^{\infty}\gamma^{(K-k)}r(k)=\max_{\alpha_{d_j},\beta_{d_j}^k,\delta_{d_j}^k,p_{j,k}}\sum_{K=k}^{\infty}\sum_{k=0}^{\infty}\gamma^{(K-k)}\mathrm{EE}(k,\pi) \quad (26)$$

得到s_k,a_k和$r(k)$后，agent与环境的交互过程可由式(27)所示的轨迹表示：

$$\mathrm{Trajectory}=s_1,a_1,r(1),s_2,a_2,r(2),\cdots,s_K,a_K,r(K) \quad (27)$$

3.2 深度强化学习算法

为了解决上述的DFMDP问题，经典的Q-Learning算法是一种有效的工具[13]. 正如之前提到的，本文的目标是为每个agent找到最优策略$\pi^*\left(s_{d_j}^k\right)\to A$，以最大化能效. Q-Learning算法可以作为候选算法来获得解决方案. 然而，当状态-动作空间较小时，Q-Learning算法可以有效地获得最优策略. 实际上，在本文的复杂模型中，空间通常很大. 因此，Q-Learning算法无法在可接受的时间内找到最优策略. 所以，本文提出了一种采用以深度Q网络(Deep Q-Network，DQN)来替代经典Q-Learning中Q表的深度强化学习(Deep Reinforcement Learning，DRL)算法来推导Q的近似值$Q(s^k,a^k)$. 因此，DQN在第k个时隙的Q值可改写为$Q(s^k,a^k,\omega)$，其中ω是深度神经网络(Deep Neural Network，DNN)的权重. 最后，最优策略$\pi^*(s)$可近似表示为

$$\pi^*(s)=\mathrm{argmax}_{a^k}Q^*(s^k,a^{k+1},\omega) \quad (28)$$

其中，$Q^*(s,a)$是通过DNN逼近所得出的最佳Q值. DQN会选择近似动作$a^{k+1}=\pi^*(s^{k+1})$. 那么近似的$\tilde{Q}\left(s^k,a^k\right)$可由式(29)给出：

$$\tilde{Q}(s^k,a^k,\omega)=r(s^k,a^k,\omega)+\gamma\cdot\max_{a^k}\left[Q(s^{k+1},a^{k+1},\omega)\right] \quad (29)$$

其中，ω的值通过最小化损失来更新：

$$L=E\left[\left(\tilde{Q}\left(s^k,a^k,\omega\right)-Q\left(s^{k+1},a^{k+1},\omega\right)\right)^2\right] \quad (30)$$

本文使用的DQN算法结构如图2所示. 它由目标值网络和当前值网络组成. 当前值网络具有最新的网络参数，计算当前状态-动作对的价值，并且定期更新目标值网络的参数，使其计算目标Q值. 经验回放储存了agent的历史行为信息，从而打破了经验池中数据相关性和非静态分布的问题. 此外，在此结构中的两个网络均采用DNN网络，该DNN包含了两个全连接的隐藏层，其中分别设置了64个神经元和32个神经元. 输出层也为一个全连接层，只有一个神经元，其激活函数为线性整流单元(Rectified Linear Unit，ReLU)，该神经网络的结构示意图如图3所示.

算法1为本文提出的基于DQN的EH-CMNs动态资源分配算法的训练程序伪代码. 值得注意的是，在训练调参过程中，考虑的主要DQN超参数有：折扣因子γ、网络结构、学习率α、缓存尺寸、探索时间占比、最终epsilon和目标值网络更新频率等. 算法2所示为获得训练好的DQN后，本文所提出的资源分配算法的伪代码.

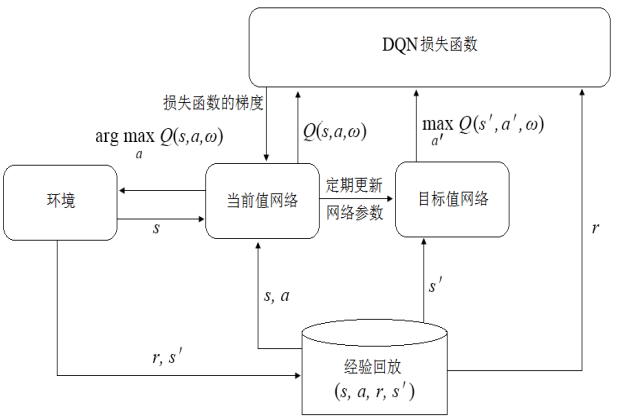


图2 DQN算法结构示意图

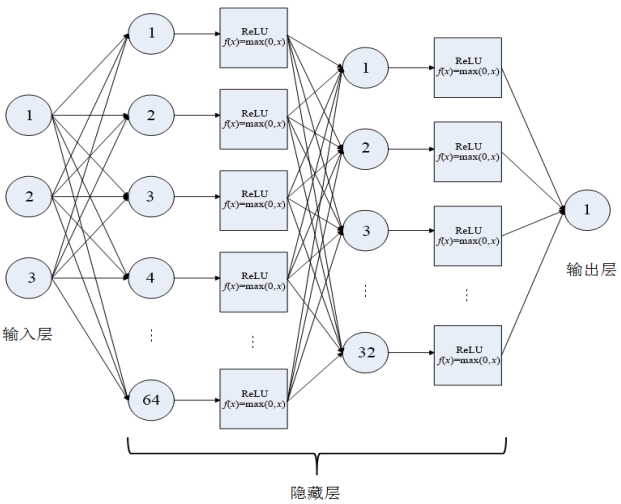


图3 神经网络的结构示意图

其中，a^k即为第k个时间步的资源分配输出，包括：传输模式、时隙分配、中继选择和传输功率. 此外，为了更清晰地反映马尔可夫决策过程中agent的状态，表1所示为第j个DU的状态空间. 由于本文中的agent在第k个时间步的状态空间仅包括$\mathrm{DQ}_{d_j}^k$和$\mathrm{EQ}_{d_j}^k$，因此状态空间是二维的.

4 仿真结果与分析

4.1 仿真设置

在仿真中，本文考虑了一个蜂窝场景下的EH-CMN，其中具备认知能力的设备和CUs随机部署在半径为800 m的蜂窝小区中，BS位于此拓扑的中心. 两个设备间的通信范围随机设置在$[20,50]$ m之间，并且为了避免严重干扰，CU和M2M对之间的最小距离设置为200 m. 此外，假设只有M2M设备配置了EH功能，并且能量收集过程是泊松分布的，到达时刻为t^k，速率为λ_e. 分组的到达过程也是泊松分布的，到达时刻为t^k，速率为λ_d. DQN事先并不知道这些先验知识. 仿真中，为每个episode设置了150个时刻，能效被求平均以减少不

算法1　所提出的DQN资源分配算法的训练过程

初始化回放缓存D为机器的数量M
初始化当前值网络的Q权重为随机的ω
for episode = 1 to U do
初始化EH-CMNs场景,并初始化状态为s_1
　for k = 1 to K do
　　以概率ε选择一个随机动作$a^k\left(\alpha_{d_j},\beta_{d_j}^k,\beta_{d_j}^k p_{j,k}\right)$
　　否则,选择$a^k=\arg\max Q^*\left(s^k,a^k,\omega\right)$
　　执行a^k,并得到即时奖励$r^k\left(\mathrm{EE}_{d_j}^k\right)$和下一状态$s^{k+1}$
　　($\mathrm{DQ}_{d_j}^{k+1}$和$\mathrm{EQ}_{d_j}^{k+1}$)
　　将$\left(s^k,a^k,r^k,s^{k+1}\right)$存储在回放缓存$D$中
　　从回放缓存D中随机采样$\left(s^k,a^k,r^k,s^{k+1}\right)$的小批量样本
　　通过随机梯度下降的ω最小化损失函数更新DNN权值
　　经过固定步长后按照下式来更新策略
　　$\pi\left(s^k\right)=\arg\max Q^*\left(s^k,a^{k+1},\omega\right)$
　end for
end for

算法2　所提出的DQN资源分配算法

加载训练好的目标值网络的参数ω
for episode = 1 to U do
　初始化EH-CMNs场景,并初始化状态为s_1
　for k = 1 to K do
　　从目标值网络接收一个动作a^k,a^k即为第k个时间步的资源分配输出:传输模式、时隙分配、中继选择和传输功率:$a^k=\left(a_{d_j}(k),b_{d_j}(k),d_{d_j}(k),p_j(k)\right)$
　　执行a^k,得到即时奖励$r^k\left(\mathrm{EE}_{d_j}^k\right)$和下一状态$s^{k+1}$即$\mathrm{DQ}_{d_j}^{k+1}$和$\mathrm{EQ}_{d_j}^{k+1}$
　end for
　计算评价指标:能量效率
end for

表1　建模的马尔可夫决策过程的状态空间表

$DQ_{d_j}^k$ \ $EQ_{d_j}^k$ → $EQ_{d_j}^{\max}$						
	(0, 0)	(1, 0)	…	…	(49, 0)	(50, 0)
↓	(0, 1)					(50, 1)
	⋮		…			⋮
	⋮			…		⋮
	(0, 49)					(50, 49)
$DQ_{d_j}^{\max}$	(0, 50)	(1, 50)			(49, 50)	(50, 50)

稳定性. DQN中使用的DNN包含2个全连接的隐藏层,其中分别设置了64个神经元和32个神经元,DNN由Tensorflow 1.13.1实现. 对于每个配置,生成了100次独立运行,并平均能效性能. 为了验证所提算法的有效性,本文以经典Q-Learning算法、仅支持直接传输方案和随机功率分配方案为基准. 表2所示为用于本仿真的详细参数.

表2　仿真参数

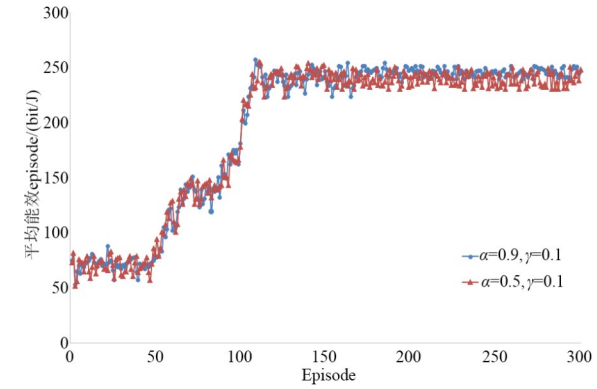
参数名称	参数值
R	800 m
设备对之间的距离	随机分布在[20, 50]
N	[1:1:30]
M	[6:2:60]
B	180 kHz
ρ_n	−174 dBm/Hz
$p_i^{\max}$	20 dBm
$p_j^{\max}$	17 dBm
λ_d	[1:1:8]分组/时隙
λ_e	[1:1:8]分组/时隙
ψ	200
τ_k	0.5 ms
$\mathrm{PS}_{\mathrm{data}}$	8比特/分组
$\mathrm{PS}_{\mathrm{energy}}$	0.0005焦/分组
$\mathrm{TR}_{\mathrm{th}}/B$	8 bps/Hz和12 bps/Hz
$\mathrm{DQ}_{d_j}^k$	50分组
$\mathrm{EQ}_{d_j}^k$	50分组

4.2　结果与分析

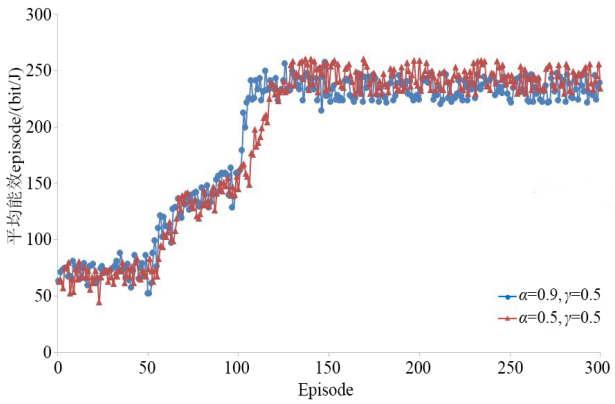
(1)学习率α和折扣因子γ对能效的影响

为了避免其他影响性能的因素,本文首先评估了学习率α和折扣因子γ对能效的影响. 假设一个场景,其中部署了一个CU和一个仅支持直接传输的M2M对,M2M对复用CU的上行频谱资源,并设置$\lambda_e=3$和$\lambda_d=5$. 图4为不同α和γ值下的平均能效. 从结果可以看出,无论是学习率α的减小还是折扣因子γ的增大,都会导致所提资源分配算法的能效趋于不稳定. 这是因为较小的α会导致较少的探索空间. 在这种情况下,所提出的算法更加关注于DNN,这将更加直接的提高用户的效能. 相反,较小的γ意味着策略优先考虑即时奖励,较大的γ会导致策略更新更具前瞻性. 因此,从长远来看,较大的γ将增加长期平均效用. 此外,从图4(c)中还可获得另一发现,虽然从长期来看,较大的γ值可以提高能效,但在与α结合的情况下,较大的α会获得较快的收敛速度,但收敛后能效波动较大,同时较小的α会导致收敛速度变慢,但能效更稳定. 另外,实验中还尝试了一些更为复杂的场景,其中部署了更多的M2M设备,但学习率α和折扣因子γ的影响是相似的. 为了

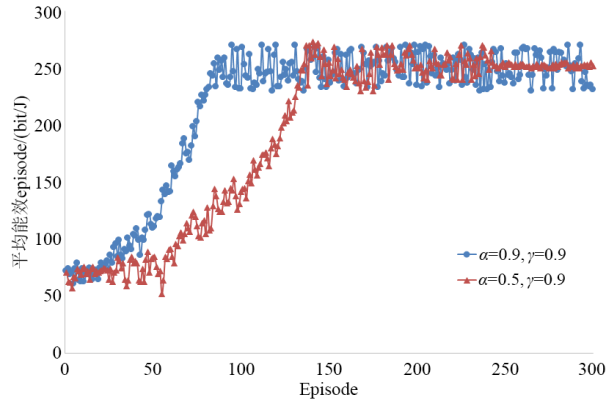
简单易懂,在此只演示了这一种情况.因此,可以做出结论,即在较高的α和较低的γ同时设置的情况下,所提算法性能更优.因此,在后面的仿真中分别设置$\alpha=0.9$和$\gamma=0.1$.



(a) $\alpha=0.9,\alpha=0.5$和$\gamma=0.1$



(b) $\alpha=0.9,\alpha=0.5$和$\gamma=0.5$



(c) $\alpha=0.9,\alpha=0.5$和$\gamma=0.9$

图4 不同α和γ值对平均能效的影响

图5所示为DRL算法和Q-Learning算法的能效优化过程.仿真结果表明了两个结论:

第一,Q-Learning算法在70episodes之前表现的要比DRL算法更好.这是因为在前70episodes中,DRL算法也要随机选择动作,并将反馈存储到回放内存中,在70episodes之后,DRL算法开始从经验中学习.值得注意的是,所提的DRL算法最初是不稳定的.然而,随着episodes的增加,表现最终趋于稳定.

第二,在此场景中,Q-Learning算法在50episodes后表现相当稳定,而不是在100episodes后,这表明Q-Learning算法比所提DRL算法具有更快的收敛速度.尽管如此,所提的DRL算法仍然在可接受时间内获得了更好的能效性能.

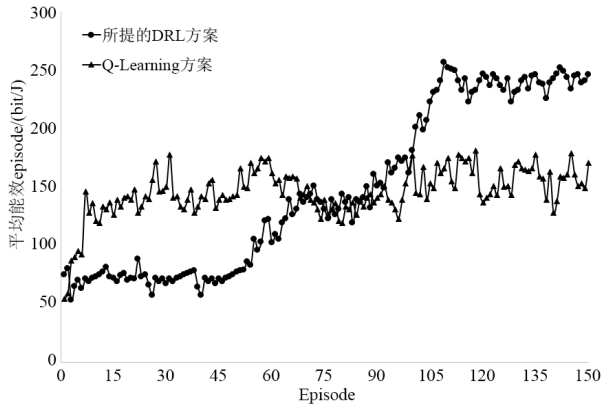
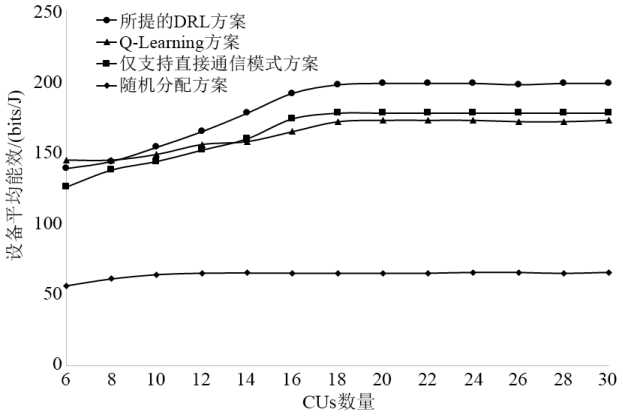


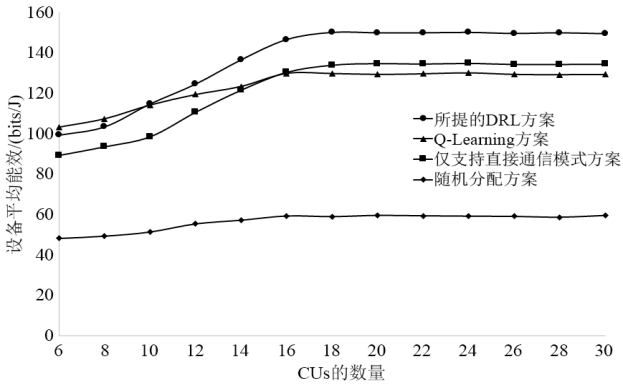
图5 平均能效的优化过程

(2)不同QoS约束下的CUs数量对能效的影响

图6为频谱效率约束分别为$TR_{th}/B=8$ bps/Hz和$TR_{th}/B=12$ bps/Hz时不同数目的CUs对能效的影响.结果表明,与仅支持直接传输方案相比,本文所提出的算法具有更高的能效.这是因为所提算法同时考虑了直接和协作传输,并通过DRL算法选择最佳传输模式.与随机功率分配方案相比,本方案综合考虑了传输模式、中继选择和分配的时隙来确定传输功率,同时发现随机功率分配方案具有最差的能效.值得注意的是,Q-Learning算法的性能最初优于DRL算法.但是,随着CUs的数量增加到10,DRL算法开始优于Q-Learning算法.这主要有两个原因:(1)当CUs数量少时,资源分配问题更简单,然而,与Q-Learning算法相比,DRL具有更高的算法复杂度,因此,能量效率更低;(2)随着CUs数量的增加的同时,DRL算法开始从经验中学习而不是重放记忆,此时性能随之上升.此外,仿真中另有一个发现是,随着CUs数量增加到14,仅支持直接传输模式的方案优于基于Q-Learning的性能.这是因为部署了更多的设备将减少两个设备之间的距离.值得注意的是,即使仅支持直接传输模式的方案也采用了DRL算法来获取相对优化的能效.此外,结合图6(a)和图6(b)发现,在较低的QoS约束条件下($TR_{th}/B=8$ bps/Hz),可以获得较高的能效.这是因为TR_{th}/B越小,p_i越小,对M2M通信的干扰越小,能效就越高.



(*a*) 频谱效率为 $TR_{th}/B=8$ bps/Hz时CUs数量对平均能效的影响



(*b*) 频谱效率为 $TR_{th}/B=12$ bps/Hz时CUs数量对平均能效的影响

图6　频谱效率约束分别为 $TR_{th}/B=8$ bps/Hz 和 $TR_{th}/B=12$ bps/Hz 时不同数目的CUs对能效的影响

(3)部署设备数量对能效的影响

图7所示为EH-CMNs中部署不同数量设备的平均能效比较. 在此仿真中,CUs的数量始终设置为10,随机部署在场景中. 仿真结果表明,DRL算法在Q-Learning算法、仅支持直接传输方案和随机功率分配方案中具有最高的能效. 最初,当设备数量较少时,所提的DRL算法、Q-Learning算法和仅支持直接传输模式方案具有相似的性能. 然而,随着设备的增多,DRL算法开始优于其他两种方案. 同时,从结果中可以发现,随着设备的增多,大多数算法(随机功率分配算法除外)的平均效能达到最高点. 当设备数量进一步增加时,能效明显降低. 出现这一现象的主要原因有4点:(1)当设备数量较少时,虽然M2M对与CU之间的干扰较小,但M2M对多为单跳模式,相对传输距离较远,因此消耗了更多的能量;(2)在设备数量较少的情况下,由于数据到达率受到 $\lambda_d=[1:1:8]$ 分组/时隙的限制,这就使得传输速率受到限制,因此能效也无法到达最大值;(3)随着设备数量的增多,将有更多的设备在DRL和Q-Learning算法的优化下,选择两跳M2M模式,在这种情况下,传输距离相对较近,因此耗能也就减少,进而提高了能效(从图7中可以看出,DRL和Q-Learning两种算法都具有相对较高的能效和以最高能效维持相对较宽的设备数量,这都源于两种算法在优化资源分配的过程中考虑了通信模式的选择);(4)当设备数量增加到一定程度后,由于可用频谱资源的有限性以及复用率的增大,导致M2M与CU之间的干扰增强,从而影响了能效. 另一个有价值的发现是,在仅支持直接传输模式方案下,设备的平均效能急剧下降,并在31 bits/J时达到最低值. 这是因为在设备数量较大时M2M对与CU之间的干扰急剧增大,能耗也会增加. 因此,可以得出一个重要结论,传输模式的选择对能效的提高有着重要的贡献.

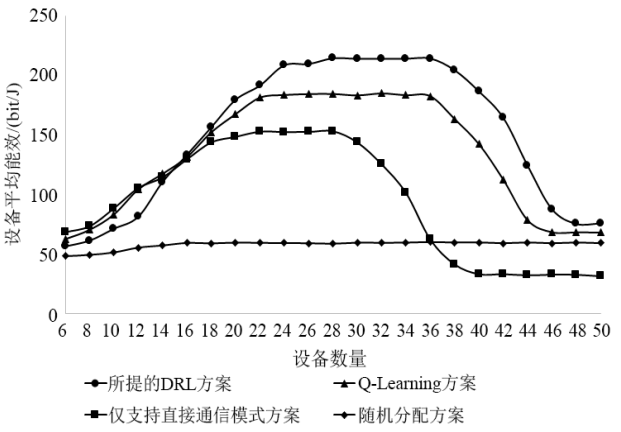


图7　不同设备数量对平均效能的影响

(4)能量收集率 λ_e 对能效的影响

图8所示为不同能量采集率 λ_e 下的能效比较. 在此仿真中,数据到达率 λ_d 设置为3以模拟物联网中M2M通信的小突发数据. 结果表明,所提的DRL算法和Q-Learning算法可以获得相对较高的能效. 随着 λ_e 的增加,能效得到了极大的提高. 这是因为 λ_e 越高,每个时隙中能够获取的能量越多. 同时,随着 λ_e 的增加,DRL算法总是具有最高的能效,这是因为它能够在能量收集时间、传输模式、中继选择和功率分配之间获得

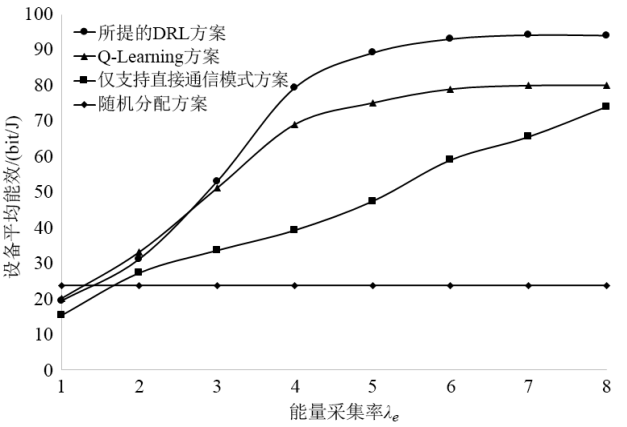


图8　能量采集率 λ_e 对平均能效的影响

最佳的相关性. 最后,发现随机资源分配方案不会改变能源效率,这是因为在分配传输功率时,它并没有考虑设备的可用能量.

5 总结

本文的主要目的是研究 EH-CMNs 的资源分配方案. 与传统 M2M 通信不同,可用能量将是资源分配中应该考虑的重要因素之一. 具体而言,本文以最大化平均能效为目标,综合考虑了每个设备的传输模式、中继选择、分配时隙、功率控制以及能量为约束,将该资源分配问题建模为一个 DFMDP. 由于该优化问题的高复杂性,提出了一种基于 DRL 的算法来求解最优解. 仿真表明,所提出的方案能够使 agent 在无须网络全局信息的情况下,自适应地从环境中学习获取资源分配的最优策略,从而显著提高能量效率.

参考文献

[1] ZHOU Z Y, GONG J, HE Y J, et al. Software defined machine-to-machine communication for smart energy management[J]. IEEE Communications Magazine, 2017, 55(10): 52-60.

[2] ZHANG C T, ZHOU Z Y, LIU P J, et al. Resource allocation for energy harvesting based cognitive machine-to-machine communications[C]//ICC 2019 - 2019 IEEE International Conference on Communications (ICC). Shanghai: IEEE, 2019: 1-6.

[3] ALHUSSIEN N, GULLIVER T A. Optimal resource allocation in cellular networks with H2H/M2M coexistence[J]. IEEE Transactions on Vehicular Technology, 2020, 69(11): 12951-12962.

[4] 蒋继胜, 朱晓荣. H2H 与 M2M 共存场景下的上行资源分配算法[J]. 电子学报, 2018, 46(5): 1259-1264.
JIANG J S, ZHU X R. An uplink resource allocation algorithm under the scenario of coexistence of H2H & M2M based on knapsack model[J]. Acta Electronica Sinica, 2018, 46(5): 1259-1264. (in Chinese)

[5] AHMAD T, CHAI R, ADNAN M, et al. Low-complexity heuristic algorithm for power allocation and access mode selection in M2M networks[J]. IEEE Internet of Things Journal, 2022, 9(2): 1095-1108.

[6] ZHOU Z Y, GUO Y F, HE Y H, et al. Access control and resource allocation for M2M communications in industrial automation[J]. IEEE Transactions on Industrial Informatics, 2019, 15(5): 3093-3103.

[7] JANG H S, PARK H S, SUNG D K. A non-orthogonal resource allocation scheme in spatial group based random access for cellular M2M communications[J]. IEEE Transactions on Vehicular Technology, 2017, 66(5): 4496-4500.

[8] NOBAR S K, AHMED M H, MORGAN Y, et al. Uplink resource allocation in energy harvesting cellular network with H2H/M2M coexistence[J]. IEEE Transactions on Wireless Communications, 2020, 19(8): 5101-5116.

[9] 田辉, 王聪, 马文峰, 等. 一种人与人和机器到机器共存下能效最大化的上行用户分配算法[J]. 电子与信息学报, 2021, 43(10): 2902-2910.
TIAN H, WANG C, MA W F, et al. A user association algorithm for maximizing energy efficiency with human-to-human and machine-to-machine coexistence[J]. Journal of Electronics & Information Technology, 2021, 43(10): 2902-2910. (in Chinese)

[10] 徐少毅, 高帅. 机器对机器通信中一种基于能量效率与系统容量的多目标无线资源管理算法[J]. 电子与信息学报, 2019, 41(12): 2817-2825.
XU S Y, GAO S. Energy efficiency and system capacity based multi-objective radio resource management in M2M communications[J]. Journal of Electronics & Information Technology, 2019, 41(12): 2817-2825. (in Chinese)

[11] MITRAN P. On optimal online policies in energy harvesting systems for compound Poisson energy arrivals[C]//2012 IEEE International Symposium on Information Theory Proceedings. Cambridge: IEEE, 2012: 960-964.

[12] BAXTER L A. Markov decision processes: Discrete stochastic dynamic programming[J]. Technometrics, 1995, 37(3): 353.

[13] CAI X J, ZHENG J, ZHANG Y. A Graph-coloring based resource allocation algorithm for D2D communication in cellular networks[C]//2015 IEEE International Conference on Communications (ICC). London: IEEE, 2015: 5429-5434.

作者简介

许艺瀚 男,1985 年 8 月生,江苏南京人. 博士,副教授,硕士生导师. 主要研究方向为人工智能、无线资源管理、物联网.
E-mail: xuyihan@njfu.edu.cn

田永波 男,1997 年 6 月生,江苏无锡人. 南京林业大学信息科学技术学院研究生. 主要研究方向为机器学习、无线资源管理和机器到机器通信.
E-mail: tianyongbo@wayzim.com