文章编号：2095-4980(2022)04-0393-09

改进 Q–Learning 的 WRSN 充电路径规划算法

刘　洋[1]，王　军[*1,2]，吴云鹏[1]

(1.苏州科技大学 电子与信息工程学院，江苏 苏州 215009；2.中国科学院长春光学精密机械与物理研究所，吉林 长春 130033)

摘　要：针对传统无线传感器网络节点能量供应有限和网络寿命短的瓶颈问题，依据无线能量传输技术领域的最新成果，提出了一种基于改进 Q–Learning 的无线可充电传感器网络的充电路径规划算法。基站根据网络内各节点能耗信息进行充电任务调度，之后对路径规划问题进行数学建模和目标约束条件设置，将移动充电车抽象为一个智能体(Agent)，确定其状态集和动作集，合理改进 ε-greedy 策略进行动作选择，并选择相关性能参数设计奖赏函数，最后通过迭代学习不断探索状态空间环境，自适应得到最优充电路径。仿真结果证明：该充电路径规划算法能够快速收敛，且与同类型经典算法相比，改进的 Q–Learning 充电算法在网络寿命、节点平均充电次数和能量利用率等方面具有一定优势。

关键词：无线传感器网络；改进 Q–Learning；充电路径规划；ε-greedy 策略；奖赏函数

中图分类号：TP393　　**文献标志码：**A　　**doi：**10.11805/TKYDA2020729

WRSN charging path planning algorithm for improved Q–Learning

LIU Yang[1]，WANG Jun[*1,2]，WU Yunpeng[1]

(1.School of Electronics and Information Engineering，Suzhou University of Science and Technology，Suzhou Jiangsu 215009，China；
2.Changchun Institute of Optics，Fine Mechanics and Physics，Chinese Academy of Science，Changchun Jilin 130033，China)

Abstract：Aiming at the bottleneck problems of traditional Wireless Sensor Network(WSN) nodes like limited energy supply and short network life，based on the latest achievements in the field of wireless energy transmission technology，a charging path planning algorithm based on improved Q–Learning Wireless Rechargeable Sensor Network(WRSN) is proposed. Firstly，the base station performs charging task scheduling based on the energy consumption information of each node in the network；and then mathematical modeling and target constraint setting are performed on the path planning problem. The mobile charging vehicle is abstracted as an agent，and its state set and action set are determined. The ε-greedy strategy is reasonably improved for action selection，and the relevant performance parameters are selected to design the reward function. Finally，the state space environment is explored through iterative learning to adaptively obtain the optimal charging path. The simulation results prove that the charging path planning algorithm can quickly converge，and has certain advantages in terms of network life，average charging times of nodes and energy utilization compared with the classic algorithms of the same type.

Keywords：Wireless Sensor Network；improved Q–Learning；charging path planning；ε-greedy strategy；reward function

无线传感器网络(WSN)作为物理与信息世界连接的接口，目前已广泛用于智慧建筑、国防安全、医疗农业和环境监测等领域。由于传感器节点普遍体积较小，本身所带电量有限，长时间维持节点可靠运行的能量问题是制约 WSN 进一步发展的关键。为解决 WSN 的能量问题，Tong 和 Li 等创造性地提出了使用移动充电车(Mobile Charging Equipment，MCE)为节点充电，通过运用磁共振耦合技术将 MCE 的能量无线传输给节点，极大延长了节点的工作寿命，如今已成为克服传统电源问题的主流方法[1]。

收稿日期：2020-12-24；**修回日期：**2021-04-02

基金项目：江苏省研究生科研创新资助项目(KYCX17_2060)；近地面探测技术重点实验室资助项目(TCGZ2018A005)

[*]**通信作者：**王　军　email:wjyhl@126.com

受无线可充电传感器网络(WRSN)各节点能耗分布不均和无线充电设备能力等因素影响，当前WRSN的研究重点是如何规划一条高效的充电路径。关于WRSN的充电算法已有很多学者做出研究：文献[2]提出了一种分簇的充电算法，通过运用K均值算法使充电设备周期性地访问所有簇的核心位置，并在每个簇内进行一对多的无线充电；文献[3]提出了一种按需充电的算法，根据节点对电量的需求程度，基站会调遣MC按照最近任务优先策略(Nearest-Job Next with Preemption，NJNP)对节点进行能量补充；文献[4]提出了一种基于收益的最大-最小蚂蚁(Max-Min Ant System，MMAS)算法，算法通过节点的能耗自定义收益，充电设备运行需要消耗能量，每次只选择部分节点进行充电，并对充电设备定期补充能量；文献[5]通过优化MC的路径，对使用单独充电车的网络进行周期性充电，并给出了一个近似最优充电路径。但上述研究都是基于理想的模型，没有充分考虑网络内移动充电设备的约束条件，以及未兼顾节点能耗与充电频率之间的关系，明显不适用于实际使用环境。

针对上述已有研究方法的不足，充分考虑MC在实际应用中的各种约束条件，在小规模WRSN中，使用单移动充电车，提出了一种基于改进Q-Learning的充电路径规划算法。其主要研究内容是确定Q函数中的状态集与动作集，设计动作选择策略和奖赏函数，并利用Q-Learning的反馈机制迭代更新Q值，构建一个关于Agent状态与动作的Q表，最后自适应选择一条奖赏值最大的最优充电路径。实验过程中将本文提出的改进Q-Learning充电算法与NJNP算法和MMAS算法进行对比实验，并对其进行了收敛性验证。

1 Q-Learning算法

1.1 Q-Learning概念

Q-Learning是强化学习中一种基于无监督学习的方法，其内容主要包括一个Agent、一组状态集(State，S)和动作集(Action，A)。Agent在环境内基于适当的策略选择动作，在某一状态下执行任一动作，环境都会反馈一个奖赏(Reward，R)，并且更新状态，通过不断训练，Agent能够自主感知未知环境，从而获得累积奖励最大的状态动作集[6]，其基本模型如图1所示。

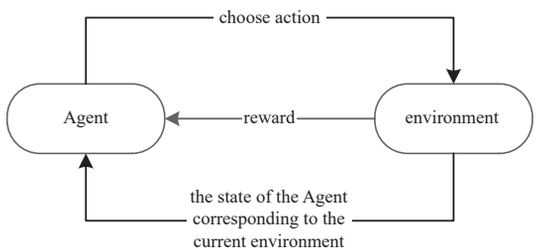


Fig.1 Diagram of Q-Learning basic model
图1 Q-Learning基本模型图

1.2 经典Q-Learning算法

针对MCE路径规划问题，将Q-Learning引入其路径决策中，经典Q-Learning开始之初，$Q(s_t,a_t)$由研究人员进行初始化设置，然后在第t个时隙选择一个动作a使状态从s_t更新至s_{t+1}，之后通过迭代计算对Q值进行更新，其更新表达式为：

$$Q_{\text{new}}(s_t,a_t)=(1-\alpha)Q(s_t,a_t)+\alpha\left[r_t+\gamma\cdot\max_{a\in A}Q(s_{t+1},a_{t+1})\right] \tag{1}$$

式中：α为学习因子；γ为奖励性衰变系数，其取值范围均为(0,1)；r_t为当前行为奖励[7]；$(1-\alpha)Q(s_t,a_t)$为旧Q值在$Q_{\text{new}}(s_t,a_t)$中所占比例；$\alpha\left[r_t+\gamma\cdot\max_{a\in A}Q(s_{t+1},a_{t+1})\right]$为本次学习得到的奖励。Agent通过遍历每个状态动作对得到节点Q值，记录并构建Q值表。

在动作集A中，用n个元素表示MCE动作集合，在状态集S中，用m个元素表示其状态集合，Q函数为一个$m\times n$的表值矩阵，它表示MCE移动或到达状态与动作趋势之间的对应关系。其一般Q-Learning算法伪代码如算法1所示：

算法1 一般Q-Learning算法伪代码

1) 初始化$Q(s, a)$；
2) 重复(算法每一步)；
3) while 当前状态≠目标状态 do；
4) 使用Q中导出的策略，从s中选择a；
5) 选择动作a_t,并观察r_t和s_{t+1}；
6) 使用式(1)更新Q值；
7) $s_t=s_{t+1}$；
8) 直到 s_t到达目标状态；
9) end while。

算法 1 描述了一般 Q–Learning 过程，首先定义一个 Q 函数 $Q(s, a)$，用以记录不同状态和动作的对应值。设置初始状态和目标状态，将 MCE 当前位置记录为当前状态，从初始位置到目标位置的运动被定义为一个完整的路径探索过程。

2　改进的 Q–Learning 路径规划算法

2.1 目标建模与约束条件

如图 2 所示，假设有 n 个静态可充电传感器节点 $\{S_1, S_2, \cdots, S_n\}$ 随机分布在边长为 L 的方形感知区域内，WRSN 由基站(Base Station，BS)、MCE、传感器节点(Sensor Node，SN)和网关节点 Gateway Node，GN)组成。

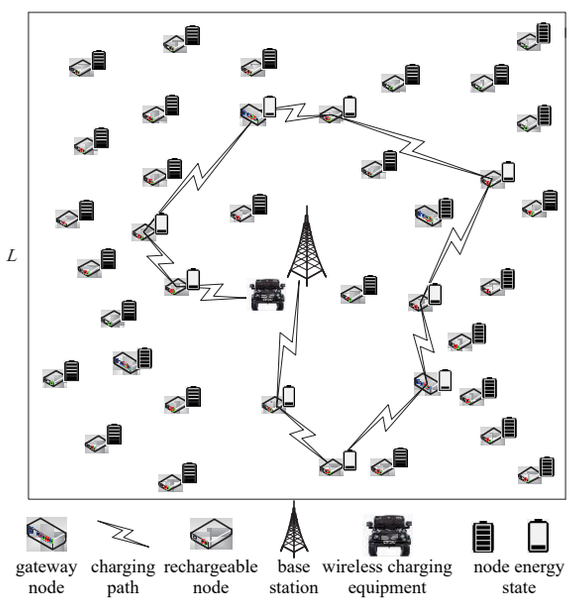


Fig.2 WRSN network model
图 2 WRSN 网络模型

MCE 在离开初始位置后，直至到达目标点才会停止，在整个过程中，MCE 的每个动作都按照一定的策略进行，并获得相应奖励，式(1)在每一步骤中更新，并且状态更新直至到达目标才会停止。当在目标位置完成充电后，以该位置为起点，进行下一段路径规划，直至最后 MCE 回到基站，完成一轮充电。

对该优化目标进行数学建模，目的是找到一条连接起始位置和目标位置的最优路径。由于 MCE 在真实环境中的状态空间是连续的，大大增加了问题求解难度，因此，首先要对规划空间进行二维离散化，从而将路径规划问题的空间搜索简化为一组离散空间节点；其中每个节点 v_k 表示 MCE 的二维坐标 (x_k, y_k)。而 MCE 路径规划问题可以描述为在空间节点中寻找几个节点，使 MCE 沿这些节点组成的路径运行总成本最小。设所有离散状态空间节点的集合为 $V=\{v_1, v_2, v_3, \cdots, v_n | v_k=(x_k, y_k)\}$。

所有充电路径的集合，包括起点和目标点，$L=\{L_1, L_2, L_3, \cdots, L_m\}$。从状态空间节点 v_i 到节点 v_j 的代价可用 $c(v_i, v_j)$ 表示，故 MCE 充电路径规划问题用数学语言描述如式(2)所示：

$$\begin{cases} C(L_k)=\sum\limits_{(v_i,v_j\in L_k)} c(v_i, v_j) \\ c(v_i, v_j)=\beta_1+e_M d(v_i, v_j) \\ \text{s.t. } L_k\in L, v_i, v_j\in V \end{cases} \tag{2}$$

式中：(v_i, v_j) 为空间节点 v_i 到节点 v_j 之间的中心距离；e_M 为 MCE 每移动单位距离所产生的能量消耗；β_1 是一个与距离相关的能耗因子，参考文献[8]给出其典型值为 2；$C\left(L_k\right)$ 表示可用路径 L_k 的总成本。考虑 MCE 携带能量 P_c 有限，充电任务要保证 MCE 有足够的能量回到 BS 进行能量补充，设定 MCE 可以运行的最大距离为 $D_{\max}$，其充电路径须满足 $\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} x_{ij}\leqslant D_{\max}$，在该二维区域内 x_{ij} 代表周期内访问节点的顺序，l_{ij} 为节点 i 到节点 j 的距离长度，用以确保 MCE 满足距离约束条件。

考虑到 MCE 移动速度对传感器节点寿命和自身能量消耗存在影响，因此在满足充电任务和距离约束的基础上，设定 MCE 一轮充电总路径长度 $L(S)=\sum\limits_{(v_i,v_j\in L_k)} d(v_i, v_j)$，单位距离能耗 $e_M=Fv$，F 为 MCE 的额定功率，对于该路径规划问题，使其满足

$$en(S+e_M L(S)\leqslant P_c \tag{3}$$

式中：$en(S)$ 是 MCE 为路径 S 上所有待充电节点充电所消耗的总能量；$e_M L(S)$ 为 MCE 路程运动所消耗的能量。MCE 移动速度设置十分重要，若速度过慢，则会影响节点的工作寿命；若速度过快，则会加剧能量消耗。充分考虑这两者关系，依据文献[9]，设定 MCE 运行速度为 2 m/s。

节点发送和接收数据时存在动态的能量损耗，因此节点传输数据量差异也会导致能量消耗变化。鉴于此，本文在充电算法中使用一种径向基函数(Radical Basis Function，RBF)预测节点动态能量消耗，其函数表达式为：

$$\Phi(P_A - X_m) = \exp\left[-\frac{1}{2\sigma} P_A - X_m \right] \tag{4}$$

式中：P_A为当前t时刻各传感器节点的剩余电量；X_m为汇聚节点的中心矢量；$P_A - X_m$代表剩余电量与汇聚节点的中心范数；σ为学习标准差，由研究人员进行初始化设定[10]。通过径向基函数进行反归一化处理，得到最终的动态预测能耗率为：$p_3 = \left[E_{ci} - \Phi(P_A - X_m) \right] / \sigma$，其中$E_{ci}$代表当前节点$i$工作状态下的能耗速率。

2.2 状态与动作空间设计

令状态集$S=\{s_1, s_2, \cdots, s_t, \cdots, s_m\}$，使用网格法离散化MCE的位置状态，将WRSN的网络区域划分成m个网格$\{s_1, s_2, \cdots, s_t, \cdots, s_m\}$，其中网格粒度大小的选取会直接影响充电路径规划算法的性能，网格粒度过大，规划路径会很不精确，但决策速度会加快；网格粒度过小，则会占用大量存储空间，但同时其有很高的位置精确度，因此本文网格粒度m的大小仅与WRSN区域大小有关。因算法是在小规模网络下执行充电任务，因此网格粒度尽可能小，使节点位置和MCE状态精确度更高[11]。在每次MCE改变位置状态后重新划分其所属状态集，以便用于Q表更新。

自由连续运动空间的动作选择，对于动作的方向和长度，一般都是无限的，因此需要模糊并简化动作的选择[12]。利用这一思想，算法设定动作集$A=\{$东,西,南,北,东北,西北,东南,西南,充电停驻$\}$，在WRSN中，一个周期代表MCE完成了一轮充电任务，MCE匀速行驶，采用改进ε-greedy策略，以$\pi(s, a)$的概率进行随机动作选择来探索环境，最终选择Q值最大的动作执行，其既可以保证MCE选取较优策略，也能保证所有状态空间都能被探索到，因此只需确定其动作方向集即可规划充电路径。

2.3 动作选择策略

Q–Learning算法通常使用ε-greedy策略来进行环境探索，在自由空间环境中，ε-greedy策略由于状态变量多，收敛速度慢，因此本文提出了一种改进的Q–Learning学习策略，其核心是应用区域分配思想，在MCE选择开始移动的动作之前，首先决定其当前位置与目标位置之间的关系。建立一个二维垂直坐标系，其中目标区域的中心位于整个自由空间的中心，坐标系将整个空间分为4个区域：

区域1：$0° \leqslant \theta < 90°$，MCE移动方向为{西，南，西南}；

区域2：$90° \leqslant \theta < 180°$，MCE移动方向为{东，南，东南}；

区域3：$180° \leqslant \theta < 270°$，MCE移动方向为{西，北，西北}；

区域4：$270° \leqslant \theta < 360°$，MCE移动方向为{东，北，东北}。

对于MCE，在每个特定区域有且只有3个可能的动作方向，因此，在寻找最佳行动时，采用区域分配的ε-greedy策略用以提高Q–Learning收敛速度。Q–Learning的最优策略通常基于状态–动作值函数$Q(s, a)$的估计，然而，在学习的初始阶段，MCE无法获得足够的环境信息，$Q(s, a)$的最优估计不够准确，往往会使得MCE与障碍物相撞，严重影响算法学习效率，因此本文设计一个评价函数$E(s, a)$来辅助状态动作值函数$Q(s, a)$选择动作[13]，其含义为MCE在某一区域状态s下分别进行3个动作，当MCE在动作后与障碍物发生碰撞时，返回−100，如果没有碰撞则返回0，其表达式如下：

$$E(s,a)_{a \in A} = \begin{cases} -100, & \text{碰撞} \\ 0, & \text{其他} \end{cases} \tag{5}$$

当$E(s, a_t)$为−100时，则将a_t从状态s的动作空间中移除。评估函数$E(s, a)$可以减少状态s下MCE的可用动作空间，避免初始阶段的大量碰撞，并且碰撞量会随着学习过程逐渐减小。当$(Q(s, a)-E(s, a))$在状态s中以概率为$1-\varepsilon$取最大值时选择动作a，或以概率ε在可用动作空间中随机选取任意动作，其表达式为：

$$\pi(s,a) = \begin{cases} \text{rand choice } a, & \varepsilon \\ \arg\max_{a \in A}(Q(s,a) - E(s,a)), & 1-\varepsilon \end{cases} \tag{6}$$

在学习初期，MCE对环境信息一无所知，因此，动作选择策略更重探索；随着学习深入，状态动作值函数可靠性越来越高，动作选择策略应发生相应改变，所以ε不是一个固定值，而是一开始较大，之后随着学习次数增加而逐渐减少的值[14]，其取值表达式为：

$$\varepsilon \leftarrow \varepsilon - \varepsilon_i (\varepsilon_{\min} < \varepsilon < \varepsilon_0) \tag{7}$$

式中：ε_0 为探索因子的初始值；$\varepsilon_{\min}$ 为探索因子最小值；ε_i 代表探索因子的增加值。

2.4 奖赏函数与参数设计

对于 MCE 充电路径规划问题，其状态空间很大，传统的稀疏奖励函数大多数动作奖励值为 0，找到有意义的动作概率相对较小，导致充电算法收敛速度较慢。为了克服稀疏奖励函数的缺点，引入更多影响充电路径性能的因素，如传感器节点剩余生命、MCE 能量消耗和目标距离，构造启发式奖励函数[15] $R\left(s_t, a_t, s_{t+1}\right)$，其表达式如式(8)所示：

$$R(s_t, a_t, s_{t+1}) = \omega_1 p_1 + \omega_2 p_2 + \omega_3 p_3 \tag{8}$$

式中：ω_1~ω_3 为各因素的影响权重，由研究人员设定；p_1~p_3 为路径性能的评估因子，其中 p_1 是 MCE 从状态 s_t 到状态 s_{t+1} 的能量消耗成本，其主要目的是为了确保 MCE 在尽可能短的时间内到达目的地；p_2 主要表示 MCE 越靠近目标点，获得的回报越多，这会为 MCE 充电的选择提供方向指导；p_3 为当前目标节点的预测动态能耗，MCE 可以依据此信息进行节点剩余生命预测，并合理规划下一充电目标节点。

为进行 Q 值迭代，需确定 Q 函数中各参数值，根据式(1)可知学习因子 α 以及奖励衰变系数 γ 是 Q–Learning 算法中最主要的参数。α 越接近 1，算法的范围搜索速度越快，但不能保证其收敛性；反之，算法足够收敛但得到的解可能只是局部最优。为平衡这两者之间的关系，令 α=0.5，使 MCE 有较为广泛且足够收敛的充电范围；当奖励衰变系数 γ 接近 1 时，算法更多考虑执行动作之后的后续收益；而接近 0 时，更考虑当前收益。本文目标是使充电效率最大化，考虑的是最终收益，因此选取较大的奖励衰变系数，取为 0.9[16]。

2.5 改进 Q–Learning 算法

在每一轮充电过程中通过迭代学习得到最优充电路径，其伪代码设计如算法 2 所示：

算法 2： 改进 Q–Learning 路径规划算法

1) Initialize：环境 V、状态集 S 和动作集 A；
2) Initialize：起始点位置和目标点位置；
3) Set up:ε,ε_0,ω,p 和最大迭代次数 N；
4) 根据网格法划分状态空间；
5) for i=0→N do；
6) s=s_0，记录起始点与目标点距离 D；
7) 步数：δ=0；
8) while (L<D) do；
9) random $\varepsilon(\varepsilon \in (0, 1))$；
10) if ε>ε_0 then；
11) 在状态 s 的动作空间中随机选择动作 a；
12) end if；
13) 得到新状态 s_{t+1}；
14) t=t+1；
15) L=D；
16) $R\left(s_t, a_t, s_{t+1}\right) = \omega_1 p_1 + \omega_2 p_2 + \omega_3 p_3$；
17) if s_{t+1} 属于指定区域 then；
18) 将参数代入式(1)；
19) else；
20) Q(s_t,a_t)=None $Q(s_t, a_t) = None$；
21) Break；
22) end if；
23) 更新 MCE 位置与目标点距离；
24) end while；
25) 更新 Q 表；
26) δ=δ+1；

27) end for。

在算法 2 中的路径规划伪代码中，δ 为学习时间步数，N 为迭代次数总数，$\varepsilon, \varepsilon_0, \varepsilon_{\min}$ 是基于区域分配的动作选择策略参数，ω 和 p 是前面讨论的奖励函数参数，D 是初始点到目标点距离，L 为 MCE 当前位置到目标点距离。

3 仿真结果与性能分析

3.1 仿真参数设置

本文通过 MATLAB 仿真软件模拟无线可充电传感器网络环境，并与 NJNP 算法和 MMAS 算法进行对比，验证 Q–Learning 充电路径规划算法的充电性能[17]。为了使实验结果充分准确，在边长为 150 m 的方形区域内随机部署 n={50,100,150,200,250,300}个可充电传感器节点，基站位于区域中心，设置 MCE 的最大运行距离 $D_{\max}$ 为 600 m，初始位置坐标为[75,75]，对于所有传感器节点随机生成初始能量，每个节点额定能量为 E_m=50 J；且各节点都配备有 GPS 模块，可以获取节点当前位置信息，当 WRSN 开始工作 1 000 s 后，MCE 开始对网络内节点进行充电，每周期充电时间间隔 250 s，周期数为 T，在进行 Q 值迭代时，设置初始学习率 α=0.5，奖励衰变系数 γ=0.9，其具体仿真参数和性能评价指标如表 1 所示：

表 1 仿真参数表

Table1 Simulation parameters

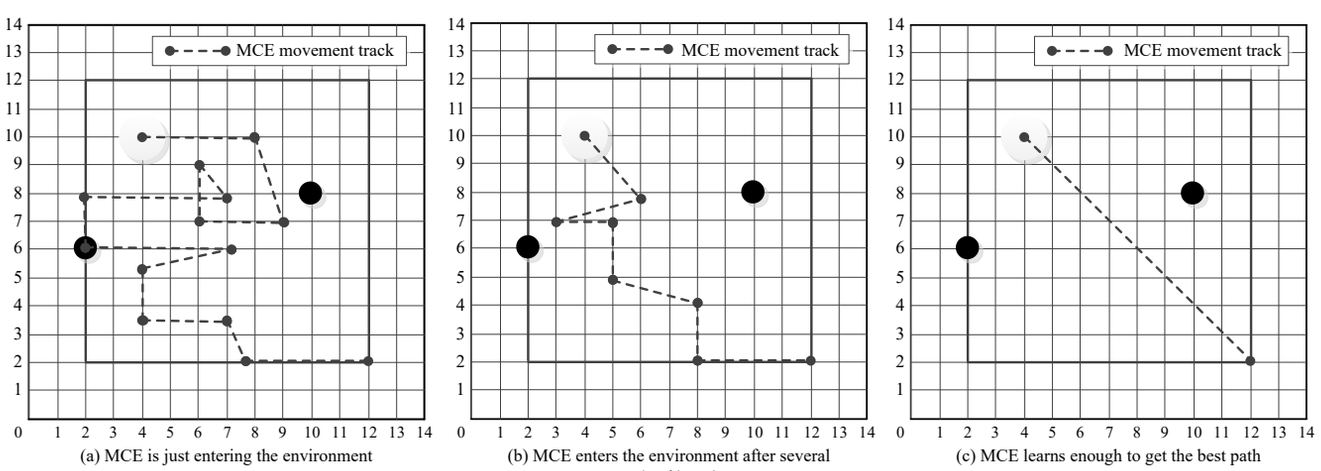
parameter definition	value	parameter definition	value
simulation area/m^2	150×150	node rated energy E_N/J	50
number of sensor nodes	50,100,⋯,300	node communication radius l/m	25
MCE driving speed v/(m·s^{-1})	2	node information packet size/bit	100
MCE battery capacity P_c/J	4 000	node sleep energy threshold E_s/J	4
MCE charging power F/(J·s^{-1})	5	node energy consumption E_{ci}/(mJ·s^{-1})	10
MCE energy consumption per unit distance e_M/(J·m^{-1})	2	number of charging cycles T	50,70,⋯,130,150

3.2 改进 Q–Learning 性能分析

使用网格法将 150 m×150 m 的仿真区域划分为 14×14 的网格部分，图 3(a)~(c)分别显示了 MCE 在所提算法最初、学习过程中和学习结束时确定的小范围空间中的移动路径。图中，黑色虚线表示 MCE 轨迹，灰色实线代表自由空间环境边界，黑色小圆形区域为障碍物，灰色圆圈表示目标区域。当 MCE 进入目标区域后，则认为其到达目标节点。

图 3(a)为 MCE 首次进入环境时的初始轨迹，在没有任何先验条件的情况下，状态空间环境对于 MCE 是未知的，MCE 不知道移动到哪里，也不知道障碍物位于哪里，因此在早期必须尝试各种路径。在探索自由空间环境的过程中，MCE 会经历许多状态，其与障碍物碰撞频率很高，并在早期阶段已在自由环境中执行许多动作。

图 3(b)中，经过多次学习过程的迭代，MCE 已初步了解了环境，知道了障碍物的位置和不同状态的奖惩反馈。但由于对环境认识不足，还未找到最佳充电路径。



(a) MCE is just entering the environment (b) MCE enters the environment after several rounds of learning (c) MCE learns enough to get the best path

Fig.3 MCE movement trajectory in small state space

图 3 小范围状态空间 MCE 移动轨迹

图 3(c)中，经过足够多的学习迭代后，MCE 已对环境有了充分了解，成功找到最佳充电路径。其累计奖励收敛过程如图 4 所示，随着迭代次数的不断增加，MCE 对 WRSN 环境逐渐了解，算法逐渐趋于收敛，在 500 次迭代计算后，MCE 能得到一条最佳的充电路径，达到快速收敛且路径最优的目的。

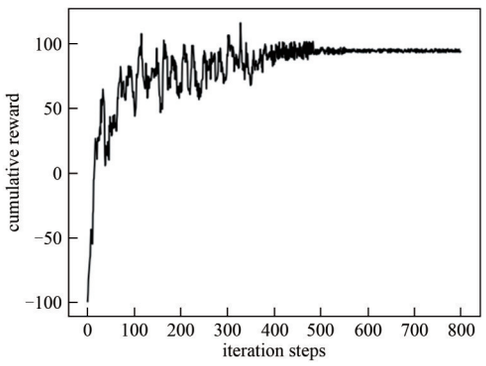


Fig.4 Cumulative reward changing with iteration steps
图 4 累计奖励随迭代步数变化曲线

3.3 充电算法性能比较

3.3.1 网络寿命

通过仿真分析 MCE 的加入对 WRSN 网络寿命的影响，在一个周期充电时间内，对不同的充电节点个数比较无充电情形、改进 Q–Learning 充电算法、MMAS 算法和 NJNP 算法在延长网络寿命方面的优劣情况，其性能比较如图 5 所示。随着传感器节点数量增加，这几种状态下的网络寿命都呈现上升趋势，并且使用了充电算法的传感器网络，相较于无充电网络明显增加了其网络寿命。对于 MMAS 算法和 NJNP 算法，不同节点数量它们的性能各有差异，但同时对网络寿命都有很大提升作用；本文提出的改进 Q–Learning 充电算法，在网络寿命性能方面提升最明显，这是因为 MCE 会根据动作奖赏值大小自适应地规划充电路径，对网络中承担任务过多、能量消耗过大的节点优先进行充电，并充分利用 MCE 自身的能量储备，尽可能多地为快要达到休眠阈值的节点充电，因此该算法大大延长了网络内高能耗节点的工作寿命，从而提升了整个网络的工作寿命。

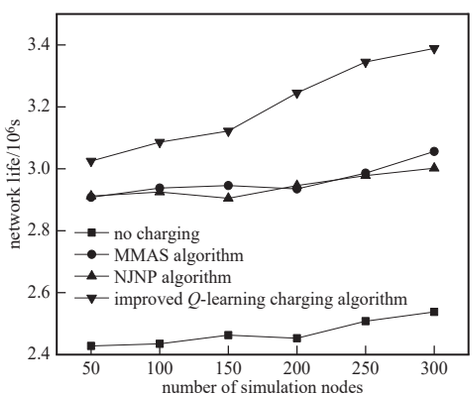


Fig.5 Network life comparison
图 5 网络寿命比较图

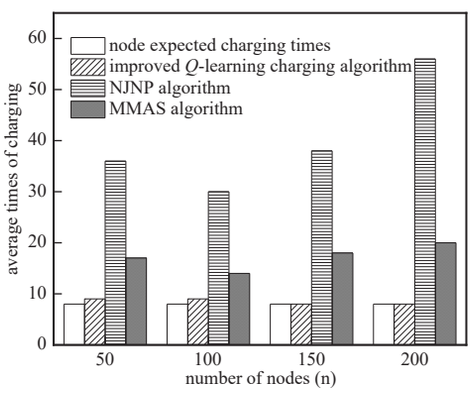


Fig.6 Comparison of average charging times
图 6 平均充电次数比较图

3.3.2 节点平均充电次数

通过比较节点的充电次数来衡量 3 种算法的优劣情况，通过改变节点的数量来验证待充电节点的选择机制。取充电周期数 T 为 30，统计不同 n 值下 NJNP 算法、MMAS 算法和本文算法的平均充电次数，并与理论求出的节点期望充电次数进行对比，其结果如图 6 所示。在同样的节点数量下，改进 Q–Learning 充电算法表现最优，在节点数量为 100 时，使用本文算法的节点充电次数与 MMAS 和 NJNP 算法相比分别减少了 76.79% 和 54.78%，并且只比节点预期的充电次数增加了 4.07%。这是因为改进 Q–Learning 充电算法以节点的能耗和剩余生命作为奖赏依据，能够更加直接地反映节点的充电需求，从而仅对距离休眠阈值较近的节点进行充电，减少了节点的充电次数。NJNP 算法根据充电任务优先策略为网络中全部节点充电，因此节点的平均充电次数远远大于期望次数；MMAS 算法是通过自定义收益的方式为部分节点进行充电，能耗较多的节点获得较多充电次数；反之，则较少。因此比 NJNP 算法的平均充电次数少，又因为节点的能耗率与充电需求并不是简单的线性关系，所以 MMAS 算法平均充电次数多于本文算法。

3.3.3 能量利用率

MCE 能量利用率也是衡量算法性能的一个重要指标，其实质为节点充电总能量占 MCE 消耗总能量的百分比。在 MCE 充电能量有限的情况下，不同的传感器节点数量会对 MCE 能量利用率产生很大的影响，其实验结果如图 7 所示。

MMAS 算法的 MCE 能量利用率最低，但随着仿真节点数的增加，3 种算法的能量利用率都有不同程度的上升，这是由于随着节点数增多，待充电节点数也会逐渐变多，MCE 需要为更多节点充电，大大增加了其充电能

量消耗，即提高了能量利用率。从图中可明显看出，本文算法的能量利用率一直高于其他两种算法，在节点数量为150个时，改进 Q–Learning 充电算法相较于NJNP算法，能量利用率提升了16.3%；MMAS算法表现不够理想；当节点数量增加至200个以上时，NJNP算法的能量利用率变得最低，这是因为在多节点时，NJNP算法的最优任务策略只是局部最优，可扩展性不高，降低了其利用率。综上所述，本文算法在MCE能量利用率方面具有优势。

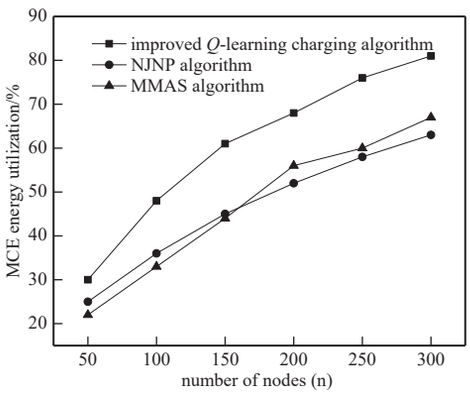


Fig.7 Energy utilization comparison
图7 能量利用率比较图

4 结论

本文针对小规模WRSN综合考虑了MCE存储能量有限、充电路径规划和节点能耗不均的情况，提出了一种基于改进 Q–Learning的无线充电路径规划算法，算法通过不断学习探索环境，在迭代计算至500步左右时，算法收敛，累计奖励趋于稳定，MCE能够自适应地规划一条最优充电路径完成任务。将该算法与经典充电算法进行仿真对比，结果表明改进 Q–Learning充电算法在网络寿命提升、节点平均充电次数和能量利用率等方面具有更好的性能。

在未来研究中会在算法中考虑无线充电设备的充电延时和大范围WRSN对其充电性能的影响，使算法可以适应更为复杂的工作场景。

参考文献：

[1] 何灏,陈永锐,易卫东,等. 无线可充电传感器网络中固定充电器的部署策略[J]. 通信学报, 2017,38(Z1):156–164. (HE Hao, CHEN Yongrui,YI Weidong,et al. Deployment strategy for static chargers in Wireless Rechargeable Sensor Network[J]. Journal of Communications, 2017,38(Z1):156–164.)

[2] LIN C,WU G W,MOHAMMAD S,et al. Clustering and splitting charging algorithms for large scaled wireless rechargeable sensor networks[J]. The Journal of Systems & Software, 2016,113(1):381–394.

[3] HE L,KONG L,GU Y,et al. Evaluating the on–demand mobile charging in Wireless Sensor Networks[J]. IEEE Transactions on Mobile Computing, 2015,14(9):1861–1875.

[4] 刘蕴娴,朱江,徐雁冰,等. 无线可充电传感器网络充电算法研究[J]. 传感器与微系统, 2020,39(2):14–17. (LIU Yunxian,ZHU Jiang,XU Yanbing,et al. Research on charging algorithm of wireless rechargeable sensor network[J]. Transducer and Microsystems Technologies, 2020,39(2):14–17.)

[5] 朱金奇,冯勇,孙华志,等. 无线可充电传感器网络中能量饥饿避免的移动充电[J]. 软件学报, 2018,29(12):3868–3885. (ZHU Jinqi,FENG Yong,SUN Huazhi,et al. Mobile charging to avoid energy starvation in wireless rechargeable sensor networks[J]. Journal of Software, 2018,29(12):3868–3885.)

[6] LI F Y,QIN J H,ZHENG W X. Distributed Q-Learning based online optimization algorithm for unit commitment and dispatch in smart grid[J]. IEEE Transactions on Cybernetics, 2019,50(9):4146–4156.

[7] YUAN Y L,YU Z L,GU Z H,et al. A novel multi-step Q–learning method to improve data efficiency for deep reinforcement learning[J]. Knowledge-Based Systems, 2019,175(1):107–117.

[8] 蒋宝庆,陈宏滨. 基于 Q 学习的无人机辅助WSN数据采集轨迹规划[J/OL]. 计算机工程, 2021,47(4):127–134. (JIANG Baoqing,CHEN Hongbing. UAV-assisted WSN data acquisition trajectory planning based on Q learning[J]. Computer Engineering, 2021,47(4):127–134.)

[9] NISIOTI E,THOMOS N. Fast Q-learning for improved finite length performance of irregular repetition slotted ALOHA[J]. IEEE Transactions on Cognitive Communications and Networking, 2020,6(2):844–857.

[10] XIAO L,XU D J,XIE C X,et al. Cloud storage defense against advanced persistent threats:a prospect theoretic study[J]. IEEE Journal on Selected Areas in Communications, 2017,35(3):534–544.

[11] DABBAGHJAMANESH M,MOEINI A,KAVOUSI–FARD A. Reinforcement learning-based load forecasting of electric vehicle charging station using Q–learning technique[J]. IEEE Transactions on Industrial Informatics, 2021,17(6):4229–4237.

[12] NAJEEB N,DETWEILER C. Extending Wireless Rechargeable Sensor Network life without full knowledge[J]. Sensors(Basel, Switzerland), 2017,17(7):200–227.

[13] 王汝言,李宏娟,李红霞. 基于SMDP的虚拟化无线传感网络资源分配策略[J]. 太赫兹科学与电子信息学报, 2020,18(1):66–

71. (WANG Ruyan,LI Hongjuan,LI Hongxia. SMDP_based resource allocation in virtualized WSN[J]. Journal of Terahertz Science and Electronic Information Technology, 2020,18(1):66-71.)

[14] LIN C,ZHOU J Z,GUO C Y. TSCA:a temporal-spatial real-time charging scheduling algorithm for on-demand architecture in Wireless Rechargeable Sensor Networks[J]. IEEE Transactions on Mobile Computing, 2018,17(1):211-224.

[15] 时丽平. WSNs中基于智能水滴的信宿路径规划算法[J]. 太赫兹科学与电子信息学报, 2020,18(4):718-722. (SHI Liping. An intelligent water droplet based path planning algorithm for WSNs[J]. Journal of Terahertz Science and Electronic Information Technology, 2020,18(4):718-722.)

[16] TONY A, HIRYANTO L. A review on energy harvesting and storage for rechargeable wireless sensor networks[J]. IOP Conference Series:Materials Science and Engineering, 2019,42(8):386-474.

[17] 顾剑,李文钧. 轻量级WSN分层协议栈的设计与实现[J]. 太赫兹科学与电子信息学报, 2018,16(2):312-316. (GU Jian,LI Wenjun. Design and implementation of lightweight layered protocol stack for WSN[J]. Journal of Terahertz Science and Electronic Information Technology, 2018,16(2):312-316.)

作者简介:

刘　洋(1996-)，男，江苏省泰州市人，在读硕士研究生，主要研究方向为无线传感器网络.email:1666356620@qq.com.

吴云鹏(1997-)，男，江苏省泰州市人，在读硕士研究生，主要研究方向为强化学习.

王　军(1979-)，男，江苏省苏州市人，副教授，硕士生导师，主要研究方向为物联网应用技术、光电信息检测. email:wjyhl@126.com.