Journal of Information Technology Management

Print ISSN: 2008-5893

Online ISSN: 2423-5059

Q-Learning Enabled Green Communication in Internet of Things

Mukesh Kumar

Ph.D. Scholar, School of Computer & Systems Sciences, Jawaharlal Nehru University, New Delhi-110067, E-mail: mukeshn7177@gmail.com

Sushil Kumar*

*Corresponding author, Assistant Professor, School of Computer & Systems Sciences, Jawaharlal Nehru University, New Delhi-110067. E-mail: skdohare@mail.jnu.ac.in

Ankita Jaiswal

Ph.D. Scholar, School of Computer & Systems Sciences, Jawaharlal Nehru University, New Delhi-110067. E-mail: ankita79_scs@jnu.ac.in

Pankaj Kumar Kashyap

Ph.D., School of Computer & Systems Sciences, Jawaharlal Nehru University, New Delhi-110067. E-mail: pankaj76_scs@jnu.ac.in

Abstract

Limited energy capacity, physical distance between two nodes and the stochastic link quality are the major parameters in the selection of routing path in the internet of things network. To alleviate the problem of stochastic link quality as channel gain reinforcement based Q-learning energy balanced routing is presented in this paper. Using above mentioned parameter an optimization problem has been formulated termed as reward or utility of network. Further, formulated optimization problem converted into Markov decision problem (MDP) and their state, value, action and reward function are described. Finally, a QRL algorithm is presented and their time complexity is analyses. To show the effectiveness of proposed QRL algorithm extensive simulation is performed in terms of convergence property, energy consumption, residual energy and reward with respect to state-of-art-algorithms.

Keywords: Energy balancing; QRL; Link Quality; Learning rate; Internet of Things

DOI: https://doi.org/ 10.22059/jitm.2022.86486 Manuscript Type: Research Paper University of Tehran, Faculty of Management

Received: December 12, 2020 Accepted: March 10, 2021

Introduction

The Internet of Things (IoT) is the network of physical objects-devices, vehicles, buildings and other items embedded with electronics, software, sensors and wireless network connectivity that enable these objects to collect and exchange data (Akyildiz IF et al. (2002), Kashyap, P. K. (2019) & Kashyap, P. K., Kumar, S., and Jaiswal, A. (2019)). Each of these smart device is uniquely identified by Internet address protocol (IP) to forward the data packet from source to destination (F. Bouabdallah et al. (2008)). The IoT has huge application in almost all the sectors of human being such as healthcare facilities, industrial organization, vehicular network, military operations, business organization and many more (Ishmanov F et al. (2011) & Ahmed AA and Mohammed Y, (2007)). These smart devices have limited battery power to perform complex computation and forward the data packet. Due to tremendous upsurge in the connected number of ubiquitous devices, there is large number of data packets travelled in the IoT network. So, there is need to choose the energy balanced routing path to forward the data packet so that lifetime of the network is improved.

In order to support various complicated application, IoT nodes have to perform the reliable operation with their limited energy, computational resources and bandwidth effectively so that it reduce the time-delay for data transmission using shortest routing path, transmission errors and ultimately improve the lifetime of the network. In this regard software define network is combined with the IoT network that separate the hardware and software control operation efficiently to cope with the mentioned challenged (J. Zhou, et al. (2016)). Therefore, dynamic routing rules for the IoT nodes provide novel data forwarding strategy, but lack in the presence of stochastic nature of channel path.

Machine learning (ML) techniques have been extensively used in the IoT network to finding the optimal route for data forwarding in the recent decade (C. Guestrin et al. (2004), Kashyap, P.K, Kumar, S. (2019), & A. Jaiswal et al. (2020)). Machine learning techniques provides learning ability to IoT network through experience, and reinforcement learning (RL) works on learning agent that improve its learning capability based on received rewards according to their taken action. By exploitation of their gained knowledge and exploration of the environment RL agent maximize its rewards (R. S. Sutton (2018)). Reinforcement learning techniques requires low computation resources with lower implementation efforts to output effective results with higher accuracy. The output of the system nearly optimal and has higher flexibility according to the changes in the environment without prior knowledge of the network. Thus, reinforcement learning and Q-learning are best suited techniques for routing approaches in the IoT network that build path with lower redundancy.

T. Hu and Y. Fei (2010) have proposed Q-learning based algorithm QELAR for the selection of next hop in the routing path. The selection of next hop depends upon the residual energy and the node density of the adjacent node, so that lifetime of the IoT network is improved by evenly

distribution of the energy. N. Javaid, O. A. Karim, A. Sher, M. Imran, A. U. H. Yasar, and M. Guizani (2003) have proposed multi-sink path selection for the data transmission using the local information such as residual energy, physical distance to update the Q-function. W. Guo, C. Yan, and T. Lu (2019) have proposed delay-aware routing algorithm for the underwater sensor networks using Q-learning. The selection of the next hop is greedy one in the residual energy and minimum propagation delay evaluated through physical distance. Whereas in (Z. Jin, Y. Ma, Y. Su, S. Li, and X. Fu (2017)), source nodes broadcast the topology information in the network, then each node simulate the residual energy, distance between them and to the destination node and feed the information to evaluating the reward of the Q-learning function. Then, it creates a virtual topology route for the data transmission and finally data are sent from intermediate node to destination node. However, proposed algorithms have limited computation for constant hop length and fail in the stochastic nature of channel state information. Also, edge length of shortest path routing in the terms of graph is dynamic, which is taken as constant in the above proposed algorithms that are not in the case of real environment.

Under these circumstances, there is need energy balanced routing algorithm based on reinforcement learning approach that include residual energy, physical distance and link quality of the channel for data transmission. The major contribution of the paper as follow:

1) Firstly, system models consist of network setting, energy consumption with residual energy model and energy balanced routing problem in the IoT network is presented to bring out their primary functions.
2) Secondly, an optimization problem is modelled according to Q-learning and Q-RL based energy balanced routing algorithm is presented. Further, time complexity of the presented algorithm is analysed.
3) Finally, Extensive simulations are presented to check the effectiveness of the presented algorithm in terms of convergence rate, energy consumption, edge length and residual energy with respect to state-of-art-algorithms.

The rest of the paper is divided into following sections. Section II described the system models used in the IoT network using graph approach. Section III explains the Q-learning based routing protocol in the IoT network. In Section IV, simulation and results are analyzed for the proposed algorithm and state-of-art-algorithms. Finally, Conclusion of this paper is presented along with future scope in the section V.

System Model

Network Setting

We consider an energy-constrained Internet of Thing network that has finite number of sensor nodes, which are randomly deployed in a given monitoring area. Each node in the network can

only communicate with the neighbouring nodes that are within its transmission range. Data transmission from one node to another takes place in synchronised time slots. Here, it is considered that each data transmission from a source node to destination takes place by using a number of intermediate nodes present along the route in the network. Each node has a single antenna, a finite battery which can be recharged periodically and works in a half-duplex mode.

The wireless connection between nodes of IoT are affected by many factors, such as residual energy of node, physical distance, channel gain etc. that makes the edge length and network state of dynamic nature in many scenarios. Here, we represent the network as a graph G = (V, E, L)with stochastic edge length, where V is the set of vertices i.e. the sensor nodes and E = (eij), such that vi, vj ϵ V, is the set of edges and L represents the probability distribution of each edge length. An edge exists between vertices vi and vj in the graph only when node j is the neighbor of node i. The nodes in the transmission range of a node constitute its neighbourhood. The length of edge (vi, vj) is denoted as l(vi, vj) and is considered as a random variable. The channel between neighbouring nodes is assumed to follow quasi-static block Rayleigh fading model and the channel gain $G_{i,j}$ between neighbouring nodes vi and vj are modelled as Markov chain. The transition probability of $G_{i,j}$ from G_1 *to* G_2 at any time instant t is given as $\text{þ}_{1,2}^{i,j} = prob\ (G_{i,j}^{t} = G_2 | G_{i,j}^{t-1} = G_1\ )$ and is unknown to the network or sensors.

Energy Consumption and residual energy Model

In IoT network, the energy is consumed for carrying out sensing, processing and communication (transmitting/receiving) activities. Out of these, data communication consumes most of the energy of a node so energy consumed for communication only is considered during routing. For simplicity, only the energy consumed for transmissions is accounted and energy spent for receiving is ignored as the idle and receiving nodes consume almost same amount of energy [20].According to the first order radio model presented in [20], for a message having b bits, the energy consumed for its transmission from v_i *to* v_j nodewith edge length l (v_i, v_j) is calculated as

$$E(b,l) = E_T(l) + E_{T_{amp}}(b,l) \tag{1}$$

$$E(b,l) = \begin{cases} b*E_T + b*\varepsilon_f*l^2 if l<l_0 \\ b*E_T + b*\varepsilon_{amp}*l^4 if l \geq l_0 \end{cases} \tag{2}$$

After data transmission residual energy $e_h^{res}(t)$ of a node at any hop, at time slot t can be evaluated as follow

$$e_h^{res}(t) = \min\{B^{max}, e_h^{res}(t-1) - E(b,l)(t)\} \tag{3}$$

Where B^{max} is the maximum battery capacity of a node, $l_0 = \sqrt{\frac{\varepsilon_f}{\varepsilon_{mp}}}$ is used for calculating the threshold distance (l_0) which in turn is used to determine the power loss model to be used i.e.

whether to use free space model or multipath fading model. Free space model is utilized when the distance between sender and receiver is less than the threshold distance otherwise multipath fading model is applied for calculating the energy consumed for transmission purposes. $E_T$is the energy requirements of transmitter and receiver circuit, $\varepsilon_f$and$\varepsilon_{amp}$are the energy consumed for amplifying transmission in order to attain a satisfactory signal to noise ratio (*SNR*) and l is the communication edge length.

Energy Balanced Routing Problem in IoT

The sensor nodes in the IoT network capture the desired data and forward this data to the destination node. Due to resource constraints in WSNs such as short communication range, limited processing potential and limited battery power, the source node, instead of communicating directly with the destination, communicates indirectly through its neighbours (multi-hop) as this leads to higher energy efficiency as compared to direct communication. In a multi-hop communication environment, routing algorithm is needed to find a communication path from source node to the destination node. Multi-hop communication overcomes the problem of energy inefficiency and short-range communication faced in direct communication but it steers imbalanced consumption of energy in the network (Khan, Tayyab, et al. (2019)) as the intermediate nodes deplete their battery faster while relaying the data of other nodes. The nodes in the vicinity of sink are the most affected ones. Therefore, the routing algorithm needs to find a path which balances the energy consumption of the nodes in the network so that all the nodes deplete their energy nearly at the same time which in turn results in increased network lifetime. A routing path *rp* in the network graph is defined as a sequence of distinct sensors in the WSN starting from source node and ending at destination node i.e. $rp = (v_1, v_2, \ldots\ldots, v_n)$such that $v_i \; and \; v_{i+1}$are adjacent vertices for $1 \leq i \; < n$andv_1=source node andv_n=destination node. The path *rp* having n sensor nodes has a length of $n - 1$.The main aim of this paper is to find an optimal routing path between a source and destination node in order to minimize the total energy consumption and transmission delay for a reliable communication.

Q-Learning Based Routing Protocol in IoT

In this section, we propose a Q-Learning based efficient routing protocol to find an optimal and reliable route from a source to destination node in order to reduce the total energy consumption and minimize the total transmission delay (based on shortest distance), which can ultimately improve the network lifetime.

Problem Modelling

The stochastic optimal routing-path finding problem is modelled as an MDP. Q-learning updating rules are used to learn an optimal policy. Here, the learning agent selects an action in order to interact with the environment (stochastic graph) to reach the next neighboring node in

the route, to get an optimal path from source to destination subject to maximize the expected reward obtained. MDP can be defined as follow:

State 'S':Each sensor node in the network and the corresponding channel gain towards its neighbor nodes is modelled as a state. The current sensor node in the routing path-finding process and current channel gain is considered as a current state.

Action 'A':All the out link neighbor nodes are considered in the action set of a state.

Transition 'P': The next state is determined by the action selection in current state.

Reward'R': Reward for a state-action pair (s, a) is calculated by using utility value which is the combination of nodes' residual energy, edge length and nodes' energy consumption and link quality.

Definition 1*- (edge length: distance between nodes):* Following formula is used to compute the edge length between any two node v_i, v_j.

$$l(v_i, v_j)= \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} \quad (4)$$

where (x_i, y_i) and (x_j, y_j) are coordinates of node v_i, and$v_j$respectively.

Definition 2- (*edge length based path):* Data packet is transferred from a source node to destination through n-hops towards the destination node. The optimal routing path based on edge length is represented as

$$L_h = \min(l_h, \dots + l_{h+n-1}) \quad (5)$$

Where l_h is the edge length between two nodes at h^{th}-hop. And, the path with minimum edge length guarantees that the transmission delay is minimum.

Definition 2- (*routing path based on energy****):***Data packet is transferred from a source node to destination through n-hops towards the destination node. The optimal routing path based on residual energy is represented as

$$E_h^{res} = \max(e_h^{res}, \dots + e_{h+n-1}^{res}) \quad (6)$$

Where e_h^{res} is the residual energy of transmitting node ath^{th}-hop.Residual energy of a node can be computed using Eq. (3).

Definition 3- (*routing path based on link quality****):*** Data packet is transferred from a source node to destination through n-hops towards the destination node. The optimal routing path based on better link quality is represented as

$$L\_Q_h = \max(l\_q_h, \dots + l\_q_{h+n-1}) \quad (7)$$

Where l_q_h is the normalized link quality at h^{th}-hopand given as

$$l\_q_h = G_{i,j}^t \cdot \frac{St}{St_{max}} \quad (8)$$

Where $St\ and\ St_{max}$ are signal strength at h^{th}-hop and maximum signal strength. Thus, the reward (utility) obtainedfor transition from $s^t(v_i^t,\ G_{i,j}^t)$ to $s^{t+1}\ (v_i^{t+1},\ G_{i,j}^{t+1})$ state after taking an action a^tat time slot t being at h^{th}-hop can be computed as

$$\mathrm{R_h^t} = \mathrm{W_1}.\,\mathrm{e_h^{res}} + \mathrm{W_2}.\,\mathrm{l\_q_h} - \mathrm{W_3}.\,\mathrm{l_h} - \mathrm{W_4}.\mathrm{E(b,l)_h} \tag{9}$$

Where W_1, W_2, W_3, and W_4 are some prescribed positive weights($\epsilon[0,1]$) parameter which reflect the importance of residual energy, link quality, edge length and energy consumption respectively in calculation of reward, and $j = 1,2,\dots J$ are number of neighbors for v_i^t. The weight parameters$W_1\ and\ W_3$are closely related to each other, such that if W_3 is set to zero, the presented model emphasize on the maximization of the residual energy in the routing path and transmission delay is ignored i.e., independent of number of intermediate hop. When $W_1$is set to zero, the presented algorithm pays more attention in reducing the transmission delay of packet and ignores residual energy of the sensor node. Thus, in both above case lifetime of the network is not optimal because of tradeoff between$W_1\ and\ W_3$. Thus, we can adjust these parameter values according to our needs.

To find an optimal routing path, the learning agent initially perceive the starting state , i.e., the source node and channel gain of the links towards its neighbors, and then selects an action using the current policy, till the agent arrives at the destination node. The state-value is updated using the temporal difference method. The updating rule for Q-learning is

$$\mathrm{Q(s^t,a^t)} = \mathrm{Q(s^t,a^t)} + \zeta(\mathrm{R_h^t} + \gamma\ \underbrace{\max}_{\mathrm{a}}\mathrm{Q(s^{t+1},a^{t+1})} - \mathrm{Q(s^t,a^t)}) \tag{10}$$

where, $\zeta \in [0,1]$ denotes learning rate, $\gamma \in [0,1]$is discount factor. Q-learning adopts ∈-greedy policy for action selection, i.e, it select optimal action having maximum state-value with probability 1-∈ and a random action with probability ∈. The main aim of the learning agent is to find an optimal policy $\pi(s^t)$ for selecting an optimal routing path. The optimal policy $\pi^*(s^t)$ denotes the state-value which is greater than other policy's state-value. The optimal routing problem can be expressed as

$$\mathrm{rp}\ =\ \underset{\{\mathrm{v_1},\mathrm{v_2},\dots\mathrm{v_n}\}}{\mathrm{argmax}}\ \lim \mathbb{E}\left[\sum_{\mathrm{h=1}}^{\mathrm{n-1}}\gamma^{\mathrm{h}}\mathrm{R_h^t}\right] \tag{11}$$

$Subject\ to -$

$\mathrm{l_h} \le \mathrm{l_h^{max}},\ \ \forall\ \mathrm{h} = 1,2,\dots \mathrm{n}-1$

$\mathrm{e_h^{res}} \ge \mathrm{e_h^{res\ (min)}},\ \ \forall\ \mathrm{h}$

$\mathrm{l\_q_h} \ge \mathrm{l\_q_h^{max}},\ \ \forall\ \mathrm{h}$

The formulated problem in (10) can be optimally solved by a Q-Learning method. The goal of proposed routing problem is to maximize the total reward over all routing paths starting at source node, such that rp ∈ φ andφ is the set of all paths in G starting from source node and ending at destination.

In routing process, each node in the network keeps a routing table which is used to select the next hop for the data transmission. The routing table contains the information about next possible nodes which can be reachable to all possible destinations in the network. This table is updated after each data transmission to store the information of node which is good for further data forwarding based on the obtained reward.

Q-Learning based Routing Algorithm

In this section, we present a learning based routing algorithm, particularly using Q-learning approach.

Algorithm- QLRA

1.Q-table Initialization
2.Initialization of path set ꝑ
3.D = v_n # destination node
4.Episode=0
5.For Episode $\leq$ Episode_{max}
6.$s^t = v_i^t,\ G_{i,j}^t$ # current state at time slot t
7.Sr= v_i^t # source node
8. ꝑ = Sr
9.While Sr$\neq v_n$
10. Action_set= neighbor nodes of Sr
11. Z = Action_set/ ꝑ
12. If Z is empty then
13. Break
14. End if
15. a^t = ∈-greedy(Sr) in Z based for s^t state # action at s^t
16. Obtain R_h^t ands^{t+1} after h^{th}-hop data transmission at s^t state #Q-table Update
17. $Q(s^t, a^t) = Q(s^t, a^t) + \zeta(R_h^t + \gamma \underbrace{max}_{a} Q(s^{t+1}, a^{t+1}) - - Q(s^t, a^t))$
18. $s^t = s^{t+1}$
19. ꝑ= ꝑ U Z
20. End While
21. Episode= Episode+1
22. End For
23. Final_route= ꝑ
24. Return Final_route

Initially, we initialize the Q-table with all zero values. After that current state $s^t$is observed. Based on the current state an action a^t is selected from the available actions at s^t (line no. 15). After executing the action, a reward and next state s^{t+1} are obtained (line no. 16). Using the

achieved reward, state-value $Q(s^t, a^t)$ is updated (line no. 17). And now next state s^{t+1} becomes current state. The algorithm converge either source node find a routing path to reach destination node or for the maximum number of episode.

Time Complexity

The time complexity of the Q-learning based routing algorithm mainly has three aspects: (1) the algorithm continues until it find destination node in the line no.9 i.e., number of intermediate node in the routing path is $(n-1)$. (2) Selecting an action (choose neighbor node) from the set of neighbors' nodes subject to maximize the expected discount reward in the network. The set of neighbors is represented in the form of $n \times n$ matrix i.e., but for the single node (state) linear search apply on the single desired row takes $O(n)$ time in line no.10 to line no.15. Thereafter line no.16 to line no. 19 requires constant time to update the Q-value (3).The algorithm runs in worst case are equal to the number of episode until convergence (line no.5). Thus, overall time complexity of the algorithm is $O\big(\ Episode_{max}(n-1)n)\big) = O\big(\ Episode_{max}(n^2)\big)$.

Results and Analysis

In this section, firstly, the convergence performance of proposed Q-Learning routing algorithm is analyzed over learning trails and link quality in terms of steps until convergence and reward (utility) respectively. Secondly, comparative analysis of the proposed algorithm against Random learning algorithm and without (w/o) learning algorithm done with respect four metrics: 1) Reward (Utility) 2) Residual energy 3) Energy consumption 4) Edge length. All these algorithms are simulated using same values of parameters for energy model and network conditions.

Simulation Environment

The simulation is carried out using MATLAB in a $100m \times 100m$ square area having 50 sensor nodes are randomly distributed. The communication range and initial energy of all the sensor nodes set to be20 m and 0.5 J respectively. The maximum bandwidth for each of the communication link in the network is 100Mbps. Without loss of generality, one source node and one destination node is selected randomly for the performance analysis of all the state-of-art-algorithms. The others simulations parameters are shown in Table 1.

Table 1. Simulation Parameters

Parameter	Value	Parameter	Value
B^{max}	$(0.5, 15)\ dBm$	Initial Energy	$O.5\ J$
Z	0.7	Γ	0.92
$Episode_{max}$	1000	∈	[0,1]
ε_f	3	ε_{amp}	0.9
St_{max}	50 dBm	l_q_h	[2,12]
W_1	0.7	W_2	0.5
W_3	0.3	W_4	0.4
E_T	0.05	n	50
$G_{i,j}$	$10\ dBm$		

Result Analysis

1. Convergence Performance over Learning Trails

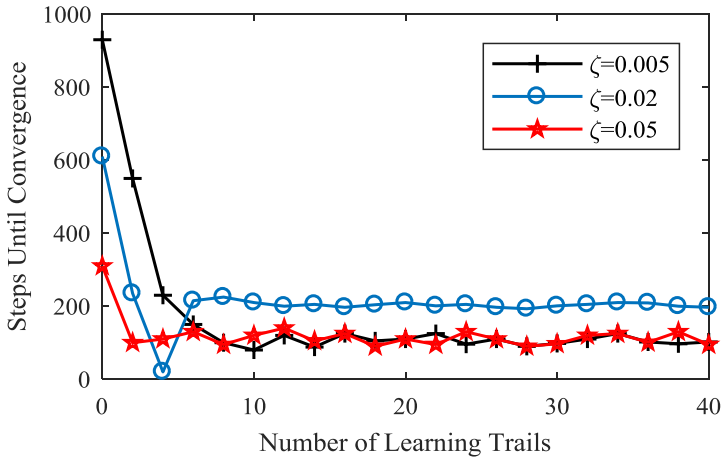

Figure 1. Convergence performance over learning trails

Figure 1 illustrates the convergence performance of the proposed QRL- based energy balanced algorithm over number of learning trails. It can be clearly observed from result, the RL agent takes 940 steps to converge for the first trails of learning rate thereafter it took less number of steps approximate 560 steps for the very less value of learning rate$\zeta = 0.005$.The number of steps on average higher than 200 steps for the taking the value of$\zeta = 0.02$, whereas the best performance to achieve convergence by the RL agent for the value of$\zeta = 0.05$. Thus, it is necessary to choose the learning rate with caution for convergence in small number of steps towards maximization of reward i.e. reduce the energy consumption and minimize the number of hop counts.

2. Reward (Utility) over Link Quality

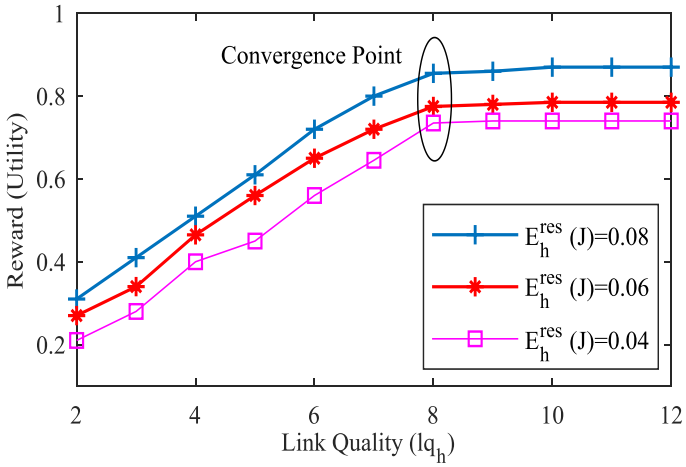

Figure 2. Reward (Utility) over link Quality

Figure 2 illustrates that the reward (utility) of the proposed Q-learning routing algorithm with respect to link quality under different residual energy of the intermediate hops in the routing path. The link quality describes the nature of communication path between the sensor nodes. And further, link quality depends upon the residual energy of the intermediate hops. If the residual energy of the communicating nodes is high then the link quality is also better and *vice-versa*. It can be observed from the results, at the beginning, the reward of the proposed algorithm increases rapidly then become stationary as the value of link quality improves. This is because of the learning capability of the proposed algorithm to optimize the reward faster at link quality's value 8. It is also worth to note down reward of the proposed algorithm cannot increase with further increase in the link quality. This is due to the fact other factors such as limited bandwidth and energy consumption in communication also increases and then affects the reward of the proposed algorithm according to Eq. (10).

3. Comparative analysis of Reward (Utility) over Episodes

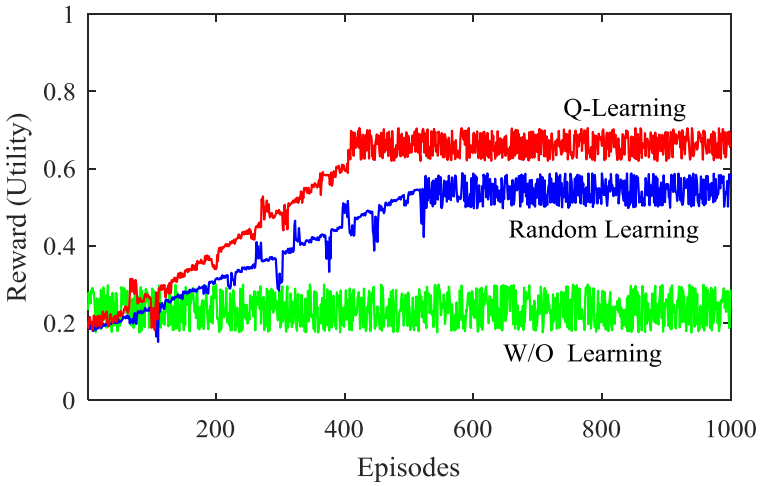

Figure 3. Reward (Utility) over Episodes

A comparison of reward (utility) between proposed QRL- based energy balanced algorithm and state-of-art-algorithm over number of episode is presented in the figure.3 using learning rate of$\zeta = 0.05$. it is clearly observed from the simulation results that Q-learning algorithm converges faster than random learning algorithm within 390 episodes. Whereas random algorithm's utility converges around 580 episodes. This is due to the fact the proposed QRL algorithm uses ϵ-greedy technique to select an action rather than randomly selected any action for the current state-reward. Also, optimal learning policy helps in the selection of an action in QRL, which ultimately maximize the reward with less number of episodes. It is also worthy to note down that worst performance is shown by without learning algorithm. This is because of neither have learning policy nor any optimization technique involved in the process of reward maximization.

4. Comparison of Residual Energy over Episodes

A comparison of convergence characteristic in the terms of residual energy between Q-learning and the state-of-art-algorithm is presented in the figure 4 using learning rate of$\zeta = 0.05$. It is clearly observed from the result as the number of episode increases residual energy increases for all the three algorithms and converges at 400 episodes. Further, it is noticeable that proposed QRL based energy balanced routing algorithm has higher residual energy about (0.87Joule) than other state-of-art-algorithm. This is because of the. QRL selects the next hop for the routing purpose based on optimal policy learning strategy subject to maximize the residual energy, better link quality and minimum distance. Whereas random algorithm select the next route based on minimum distance and does not consider residual energy of the next hop that in turns increases the overall energy consumption. And, the without learning based algorithm selects any hop for the routing randomly without considering the residual energy and minimum distance.

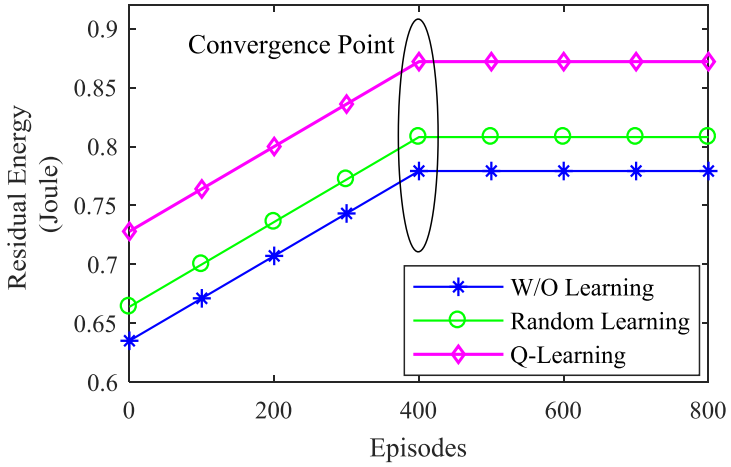

Figure 4. Residual energy over Episodes

5. Comparison of Energy Consumption over Episodes

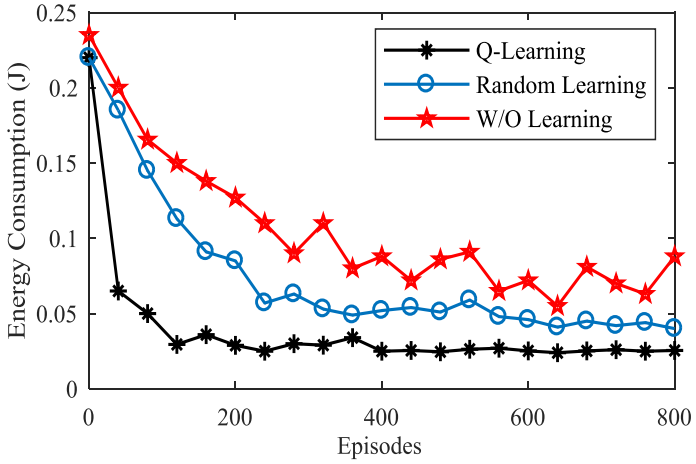

Figure 5. Energy Consumption over episodes

A comparison of energy consumption of the proposed Q-learning routing algorithm with state-of-art-algorithms over number of learning episodes is shown in the figure 5. It can be observed from the results; at the start of learning episodes the energy consumption of the proposed algorithm is 0.225 J and as the algorithm reached to 160 episodes it lower down the energy consumption at 0.03 J. Further, proposed algorithm reached up to 400 episodes, the energy reduces to 0.025 J and consumption becomes stable. Whereas other state-of-art algorithm fails to optimize the energy consumption of the nodes involved during routing path. This is because of Q-learning algorithm uses $\in$-greedy approach for the selection of optimal policy, whereas Random algorithm selects any action randomly to obtain the reward. It is also noted down that the worst performance is shown by without (w/o) learning algorithm, because it does not have any learning policy and compute reward on the current situation of node's parameters.

6. Comparison of Edge Length over Episodes

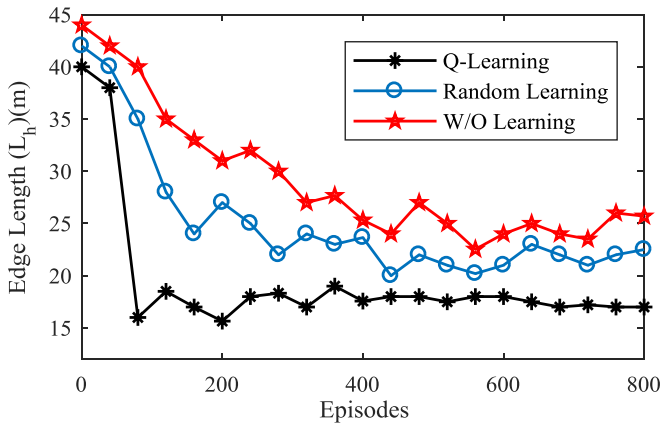

Figure 6. Edge Length over episodes

A comparison of edge length of the proposed Q-learning routing algorithm with state-of-art-algorithms over number of learning episodes is shown in the figure 6. The edge length describe the distance between the intermediate hops, smaller the edge length (*Euclidean distance*) corresponds to reduce the transmission delay and improves the also convergence speed of the learning based routing algorithm. The results shows that as the number of episode increases, the edge length of proposed algorithm reduces and stabilized about to 17 m within 400 episodes. Whereas edge length of random and without learning algorithms are fluctuates and fails to convergence. This is because of proposed Q-Learning routing algorithm at the initialization know the number of hop counts and also in learning phase $\in$-greedy policy helps to compute less number of intermediate hops count and then compute the shortest distance edge length. It can be also observed that without learning based routing algorithm compute the edge length only on *Euclidean distance* formula according to Eq. (4) and nothing to do with learning and in turn fail to converge the edge length.

Conclusion

In this paper, we handled the problem of energy balanced routing algorithm using reinforcement learning and proposed QRL algorithm for wireless sensor network. The link quality, residual energy, and distance between the two consecutive hops are used as parameter for selection of an optimal action subject to maximize the reward (utility). To achieve the objective QRL based energy balanced algorithm has been proposed and their time complexity is also analyzed to show the effectives of the proposed algorithm. It is also proved from the simulation results the proposed QRL algorithm converges faster than other state-of-art-algorithms. It is also notable from the simulation results that energy consumption and link quality and residual energy also improved compared to random algorithm and without learning algorithm. In the future, we also include the node density as another parameter to estimate the energy balanced routing path using deep learning techniques.

Conflict of interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

A. Jaiswal, S. Kumar, O. Kaiwartya, N. Kumar, H. Song and J. Lloret, (2020) Secrecy Rate Maximization in Virtual-MIMO Enabled SWIPT for 5G Centric IoT Applications," in IEEE Systems Journal, doi: 10.1109/JSYST.2020.3036417.

Ahmed AA and Mohammed Y, (2007) A survey on clustering algorithms for wireless sensor networks, *Elsevier, ComputerCommunications*, Vol. 30, pp. 2826-2841.

Akyildiz IF, Su W, Sankarasubramaniam Y and Cayirci E, (2002) Wireless sensor networks: a survey, *Computer Networks*, Vol.38, No. 4, pp. 393-422.

C. Guestrin, P. Bodik, R. Thibaux, M. Paskin, and S. Madden, (2004) Distributedregression: An efficient framework for modeling sensor network data in*Proc. 3rd Int. Symp. Inf. Process. Sensor Netw.*, pp. 1–10.

F. Bouabdallah, N. Bouabdallah, and R. Boutaba, (2008) Towards reliable and efficient reporting in wireless sensor networks, IEEE Trans. MobileComput., vol. 7, no. 8, pp. 978–994

Ishmanov F, Malik AS and Kim SW, (2011) Energy consumption balancing (ECB) issues and mechanisms in wireless sensor networks (WSNs): A comprehensive overview, *European Transactions on Telecommunications*, Vol. 22, pp. 151-167.

J. Zhou, H. Jiang, J. Wu, L. Wu, C. Zhu, and W. Li, (2016)‘ SDN-based application framework for wireless sensor and actor networks,’’ *IEEE Access*,vol. 4, pp. 1583–1594.

Kashyap, P. K., Kumar, S., and Jaiswal, A. (2019) Deep Learning Based Offloading Scheme for IoT Networks Towards Green Computing. IEEE International Conference on Industrial Internet (ICII), pp. 22-27, Orlando, FL, USA, 2019.

Kashyap, P. K., Kumar, S., Dohare, U., Kumar, V., &Kharel, R. (2019) Green Computing in Sensors-Enabled Internet of Things: Neuro Fuzzy Logic-Based Load Balancing. MDPI Electronics, 8(4), pp. 384-405.

Kashyap, P.K, Kumar, S. (2019) “Genetic-fuzzy based load balanced protocol for WSNs” International Journal of Electrical and Computer Engineering, Vol. 9, No.2, pp.1168-1183.

Khan, Tayyab, Karan Singh, Mohamed Abdel-Basset, Hoang Viet Long, Satya P. Singh, and Manisha Manjul. (2019) A novel and comprehensive trust estimation clustering based approach for large scale wireless sensor networks." IEEE Access vol. 7 pp-58221-58240.

N. Javaid, O. A. Karim, A. Sher, M. Imran, A. U. H. Yasar, and M. Guizani,(2003) Q-learning for energy balancing and avoiding the void hole routing protocol in underwater sensor networks,’’ in *Proc. 14th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Jun. 2018, pp. 702–706

R. S. Sutton and A. G. Barto, (2018) *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press

T. Hu and Y. Fei, (2010) QELAR: A machine-learning-based adaptive routing protocol for energy-efficient and lifetime-extended underwater sensornetworks,’’ *IEEE Trans. Mobile Comput.*, vol. 9, no. 6, pp. 796–809.

W. Guo, C. Yan, and T. Lu,(2019) Optimizing the lifetime of wireless sensor networks via reinforcement-learning-based routing,” *Int. J. Distrib. Sensor Netw.*, vol. 15, no. 2.

Z. Jin, Y. Ma, Y. Su, S. Li, and X. Fu, (2017) A Q-learning-based delay-aware routing algorithm to extend the lifetime of underwater sensor networks,” *Sensors*, vol. 17, no. 7

Bibliographic information of this paper for citing:

Kumar, Mukesh, Kumar Sushil, Jaiswal Ankita & Kashyap, Pankaj K. (2022). Q-Learning Enabled Green Communication in Internet of Things. *Journal of Information Technology Management*, Special Issue. 103-117.