Q-learning Based Routing Optimization Algorithm for Underwater Sensor Networks

Jia Gao, *Graduate Student Member, IEEE*, Jingjing Wang, *Member, IEEE*, Jianlei Gu, *Graduate Student Member, IEEE*, Wei Shi

***Abstract*—Underwater wireless sensor network (UWSN) plays a vital role in the field of ocean development and exploration. Designing a routing protocol for UWSN is a great challenge due to the characteristics of short lifetime and high delay. This paper proposes a Q-learning based routing optimization algorithm for UWSN. Two reward functions are designed based on the average residual energy of network, integrating factors such as energy information, transmission delay and link success rate to better balance transmission quality and lifetime. In addition, a holding time mechanism for packet forwarding is developed according to the priority of nodes. The simulation results show that compared to DBR and QLFR algorithms, this algorithm can effectively reduce transmission delay and prolong network lifetime.**

***Index Terms*—Q-learning, reinforcement learning, routing selection, underwater wireless sensor network.**

I. Introduction

UWSN is essential in promoting marine scientific research, resource development, environmental protection and security defense. UWSN is a self-organized network composed of sensor nodes with certain sensing, computing, storage and communication capabilities, autonomous underwater vehicle (AUV) and base stations. UWSN completes underwater data collection and transmission through nodes' collaboration [1], [2]. Among many research issues in UWSN area, delivering packets from a source to a destination, namely routing, is one of the fundamental problems that need to be studied for improving UWSN's performance. In a multi-hop UWSN, as shown in Fig. 1, nodes collect relevant underwater data within their sensing range, such as temperature, salinity, pressure, and other information. Subsequently, the data is transmitted to the surface base station through a series of relay nodes. Finally, the base station transmits the data to the onshore monitoring center via satellite for further analysis and processing.

Compared to traditional terrestrial sensor networks, multipath effect and doppler effect in underwater environment lead to significant signal attenuation. And there are abundant noises in underwater environment, including various types of natural and artificial disturbances such as ship noise, biological noise and thermal noise. These factors affect the reliability of underwater data transmission. Moreover, the speed of sound in water is 1500m/s, and the propagation delay of UWSN is much higher than that of land-based networks. In addition, underwater nodes are powered by batteries with limited energy. It is difficult to replace the batteries and replenish energy in complex underwater environments. Therefore, designing and optimizing of routing protocols in UWSN is challenging due to above unique characteristics [3]–[13].

Therefore, designing effective routing mechanism in UWSN is crucial for quick and energy-saving transmission. In response to issues of long transmission latency and short lifetime, this article proposes a Q-learning based routing optimization algorithm for UWSN. The main contributions of this article can be summarized as follows:

(1) This article proposes a Q-learning based routing model to balance network load and minimize transmission delay. Based on the average residual energy of network, the algorithm combines delay and energy factors to construct two different reward functions, and updates the Q-values according to the current reward dynamically.

(2) We design a timer-based candidate set coordination approach to schedule packet forwarding, where a novel holding time method is proposed on the basis of nodes' priority which reduces the transmission collision and redundant communication.

The subsequent sections of this article are structured as follows: Section II discusses related work, Section III introduces the related knowledge about underwater networks and reinforcement learning (RL). Section IV establishes and analyzes routing optimization model based on Q-learning and holding time mechanism. Section V gives the detailed design of this algorithm. And section VI introduces the experimental results in simulation. Finally, section VII summarizes our work.

II. Related Work

Numerous UWSN routing algorithms have been documented in the literature. For example, EULC [14] is a cluster-based routing algorithm. Distance and residual energy factors are considered in routing selection process. This algorithm also increases lifetime by adjusting contention radii of cluster headers (CHs) and reducing cluster scale near the sink. P-AUV [15] is a novel routing protocol designed for AUV-based networks, which aims to achieve higher energy efficiency and packet delivery ratio by the means of minimizing collisions. The proposed protocol utilizes nodes' locations for forwarding decisions and creates loop-free paths in which nodes near the base station have higher priority to be chosen as relay. In [16],

This work was supported by the National Natural Science Foundation of China (NSFC) (No.62171246, U1806201, 62101298) and the Natural Science Foundation of Shandong (ZR2021ZD12, ZR2021MF130).

Jia Gao, Jingjing Wang, Jianlei Gu and Wei Shi are with the School of information science and technology, Qingdao University of Science and Technology, Qingdao 266061, China (e-mail: gaojia1997@126.com; kathy1003@163.com; gujianlei2023@163.com; shiwei6670595@126.com).

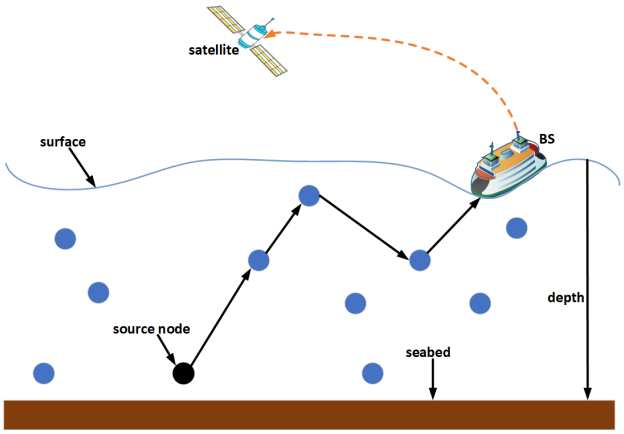

Fig. 1. Schematic diagram of multi-hop UWSN routing selection.

author proposed a DS evidence theory-based opportunistic routing scheme to enable energy efficient forwarding.

RL [17] obtains experiential information through the interaction between agent and unknown environment. It has the potential to improve the routing process under various objectives. Basagni [18] introduced a routing protocol for energy-harvesting UWSN that takes into account residual energy, foreseeable harvestable energy and link quality. This algorithm can effectively reduce packet drop ratio and improve communication reliability, but it cannot extend network lifetime. In [19], author proposed a distributed routing protocol for UWSN in order to enhance network performance. By integrating a game-theory approach with Q-learning, this algorithm tries to maximize their profit in routing decision with other agents. In order to extend network lifetime, Zhou [20] proposed a QLFR algorithm which employs a holding time mechanism. In this mechanism, nodes with lower priority dropping repeated packet, which reduces redundant transmission. The work by Valerio [21] introduced a CARMA algorithm which selects a list of relay nodes according to the local channel quality and energy consumption. This algorithm can increase the packet delivery ratio and ensure a lower energy cost. In [22], Jin investigated the congestion control problem in UWSN routing and put forward an RCAR protocol to minimize energy consumption. Firstly, this algorithm calculates pipe radius based on average residual energy of the neighbor nodes. Then, virtual routing is executed using the RL-based algorithm to select the forwarding node. QL-EDR [23] utilizes Q-learning algorithm to select next-hop based on the residual energy and distance in cluster network. In [24], author constructed a DMARL scheme which is specifically designed for underwater optical communication. The global reward is provided to gather feedback on environmental changes and this reward depends on the normalized residual energy and link quality. Moreover, this algorithm can adapt the learning rate according to the dynamic environment which accelerates the algorithm' s convergence. Lu [25] introduced an EDORQ algorithm, the reward function in this algorithm is designed with residual-energy related values, depth related values and void detection factors. RLOR algorithm [26] demonstrates enhanced performance in dense network by reducing the risk of routing void problem. In [27], author proposed a Q-learning based UWSN multi-hop cooperative routing protocol. This protocol can utilize distance information to select the node with the maximum Q value as a forwarder. Additionally, it combines collaborative communication with the Q-learning algorithm to reduce network energy consumption effectively.

Although the above mentioned algorithms can improve the routing performance of wireless sensor network, they are unable to effectively balance the relationship between network lifetime and communication quality. Moreover, many routing algorithms involve redundant transmissions, resulting in additional energy consumption and communication conflicts.

III. BACKGROUND

A. Energy Consumption Model for Underwater Acoustic Channel

In this paper, Thorp model [28] is adpoted to describe the underwater acoustic propagation. The path loss is shown as (1),

$$A(d, f) = d^k \alpha(f)^d, \tag{1}$$

of which f is the transmitting frequency, d is the transmission distance, k is the spreading coefficient.

In (1), $\alpha(f)$ is the absorption coefficient, the calculation equation is:

$$\alpha(f) = 0.11 \times \frac{f^2}{1+f^2} + 44 \times \frac{f^2}{4100+f^2} + 2.75 \times 10^{-4} f^2 + 0.003. \tag{2}$$

The energy consumed by node to send data of c-bit is,

$$E_{Tx}(k, d) = cP_0 A(d, f) = cP_0 d^k \alpha(f)^d. \tag{3}$$

Similarity, the energy consumed by node to receive data of c-bit is,

$$E_{Rx}(k) = cP_r, \tag{4}$$

where P_0 represents the minimum power required for node to send data and P_r represents the reception coefficient.

In this algorithm, receiving energy consumption includes a portion of overhearing energy consumption, and overhearing energy consumption is much lower than the energy consumption for transmitting and receiving. In terms of the overall energy consumption, the proportion of overhearing energy consumption is small and can be neglected.

B. Node Motion Model

In this paper, underwater sensor nodes are fixed by anchor chains and float in water with restricted movement in a specific area. Adjusting the length of the anchor chains can form a three-dimensional network [29]. In order to determine the range of nodes' motion, it is necessary to conduct force analysis on the nodes. Because the gravity of node is much smaller than the buoyancy, tension and water impact force, the gravity can be ignored, as shown in fig. 2, the node is subjected to transverse impact force F of water flow, buoyancy force B generated by current, and tension T of the anchor chain to the

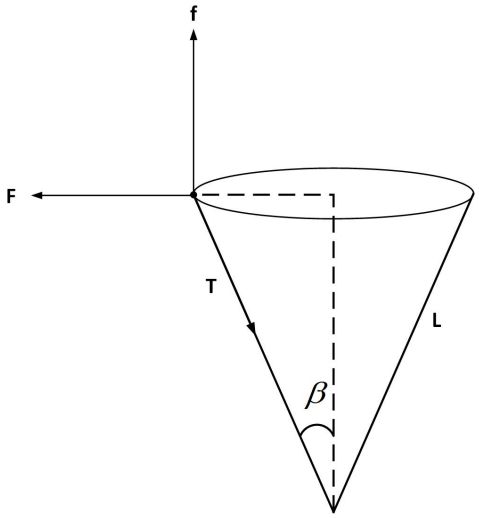

Fig. 2. Stress analysis diagram of an underwater node

node [30]. Essentially, the three forces constitute a group of balance forces, the included angle of anchor chain offset is β and the length of anchor chain is L, where $\tan\beta = F/B$, the motion range of the node is $[L\cos\beta,\ L]$.

The Random Waypoint Model is used to simulate the mobility of underwater nodes, which is described as follows: each node randomly selects a destination within a certain range and a speed that is uniformed distributed between $[v_{min}, v_{max}]$. Then, the node travels toward the newly destination at the selected speed. Upon arrival, the node pauses for $t_{pause} \in [t_{min}, t_{max}]$ before starting process again [31].

C. Reinforcement Learning

RL is a branch of machine learning which utilizes the principle of Markov decision process to learn the interaction between agent and environment. The four elements of RL are expressed as $< S, A, P, R >$, where S is the state space, A is the action space, P is the environment, and R is the immediate reward. Fig. 3 shows the interaction process between agent and environment, the agent selects an action a according to current state s and executes it firstly. When transferring from state s to next state s', the agent calculates the reward r according to the feedback of the environment with minimal a priori knowledge. The agent interacts with the environment continuously and adjusts its action strategy.

Q-learning is a algorithm based on value iteration in RL. It needs to maintain and update a Q-table with the size of $|S| \times |A|$, where $|S|$ is the number of state and $|A|$ is the number of action. The elements in Q-table represent the expected cumulative rewards generated by the agent's actions a_i in its current state s_i, that is $Q(s_i, a_i) = E_\pi[\sum_{k=0}^{+\infty} \gamma^k r_{t+k+1} | s_t = s_i, a_t = a_i]$. Q-table update formula is as follows:

$$Q_{t+1}(s_t, a_t) = (1-\alpha_t)Q(s_t, a_t) + \alpha_t[r(s_t, a_t) + \gamma max_{a\in A} Q_t(s_{t+1}, a)], \quad (5)$$

of which, α_t is the learning rate and γ is the discount rate. In continuous interaction with the environment, Q-table gradually tends to accumulate rewards in reality, and agents also tend to make optimal decisions.

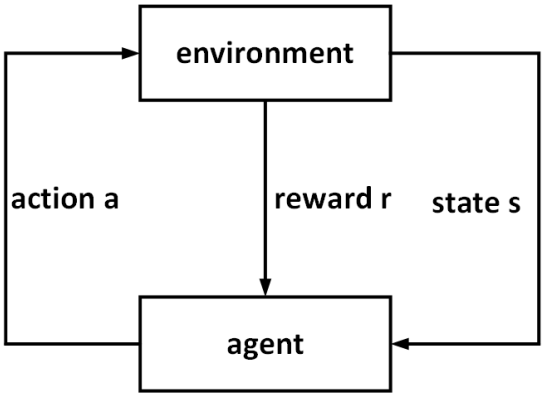

Fig. 3. Reinforcement learning model framework

IV. Model Establishment

A. Q-learning Based Routing Optimization Model Establishment

This article proposes a Q-learning based routing model for UWSN, and its structural framework is shown in Fig. 4.

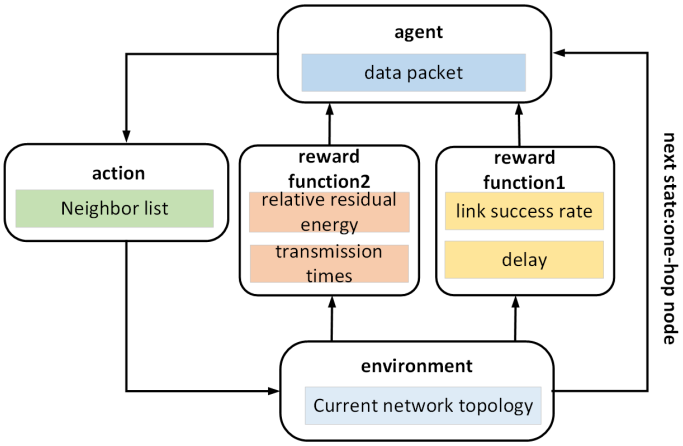

Fig. 4. Structural framework of Q-learning based routing optimization model

Agent: Each data packet in our network is regarded as an agent in our model.

Environment: The current network topology is regarded as environment.

State: The state space of this model is defined as all nodes in network and each data packet is transmitted by a series of relay nodes (states) to base station.

Action: Action is a key factor on RL, and the action set of each state is composed of its available neighbor nodes.

Reward: As the average residual energy of the network decreases, the purpose of routing selection shifts from ensuring the quality of underwater acoustic communication to balancing network load and extending network lifetime. Therefore, we constructed two reward functions. When the average residual energy of network exceeds a specific threshold, the network energy is sufficient, so paths with smaller delay and higher

transmission success rate are selected to improve communication quality. The specific reward function is shown as follows:

$$r(s_t, a_t) = \begin{cases} \frac{success(s_t, s_{t+1})}{len/C + d(s_t, s_{t+1})/v_{sound}}, & s_{t+1} \textit{ is relay node}, \\ r_1, & s_{t+1} \textit{ is target node}, \\ \text{-}r_2, & s_{t+1} \textit{ has no next hop}, \\ \text{-}r_3, & s_{t+1} \textit{ has insufficient energy}. \end{cases} \quad (6)$$

When the next state s_{t+1} is a relay node, the reward given is related to the link transmission success rate $success(s_t, s_{t+1})$ and delay information $len/C + d(s_t, s_{t+1})/v_{sound}$, of which len is the length of the packet, C is the channel capacity, $d(s_t, s_{t+1})$ is the distance of current node and next-hop and v_{sound} is the speed of sound in water. When the next state is the target node, we give it a positive reward r_1. But if the next hop doesn't have sufficient energy to complete the transmission task or next-hop has no neighbor nodes, a negative reward $-r_2$ and $-r_3$ are given respectively.

When the average residual energy of network belows a specific threshold, reducing and balancing node energy consumption become an important factor in network routing selection. The related reward function is:

$$r(s_t, a_t) = \begin{cases} \frac{E_r(s_{t+1}) \times O_{total}}{E_{init}(s_{t+1}) \times O_{s_{t+1}}}, & s_{t+1} \textit{ is relay node}, \\ r_1, & s_{t+1} \textit{ is target node}, \\ \text{-}r_2, & s_{t+1} \textit{ has no next hop}, \\ \text{-}r_3, & s_{t+1} \textit{ has insufficient energy}. \end{cases} \quad (7)$$

If s_{t+1} is a relay node, the reward obtained is related to factors such as energy and forwarding times, where $\frac{E_r(s_{t+1})}{E_{init}(s_{t+1})}$ is the ratio of residual energy, $O_{s_{t+1}}$ is the number of times the node s_{t+1} participates in data forwarding, and O_{total} is total number of data forwarding by each node in the network.

When the agent selects the next hop, the Q-value update formula is as follows:

$$Q_{t+1}(s_t, a_t) = (1 - \alpha_t) Q(s_t, a_t) + \alpha_t [r(s_t, a_t) + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]. \quad (8)$$

B. Holding Time Mechanism Model

If all eligible neighbor nodes take part in forwarding the same packet, it will increase communication conflicts and energy consumption. To reduce energy waste, we design a holding time for each candidate node with the principle that nodes with larger priority have shorter holding time.

Firstly, we sort neighbor nodes in descending order according to Q-value. Next, each node creates a priority list based on sorted results of neighbor nodes. When a node receives a packet, it will search for the priority list. If the node's ID is in priority list, it will hold the package for a specific period of time, called holding time. If not, it will simply drop the package. The holding time can be expressed as follows:

$$Time = \frac{\alpha}{1 + e^{-(n-1)}} \times Time_{max}, \quad (9)$$

of which α is positive parameter, n is the position of node in priority list and $Time_{max}$ is a pre-defined maximum holding time. Holding time should be long enough to suppress the duplicate transmission of lower priority nodes before transmitting the packet. Considering the worst condition, $Time_{max}$ can be defined as (10), where d_{max} denotes the maximal distance of one-hop in UWSN.

$$Time_{max} = \frac{2 \times d_{max}}{R}. \quad (10)$$

The candidate forwarder with a larger Q-value has a shorter holding time, which means that it is preferential to transmit the data packet. If a candidate node with a lower priority overhears the same packet transmission of any node within its holding time, it will drop the packet. An example of the holding mechanism is illustrated in Fig. 5.

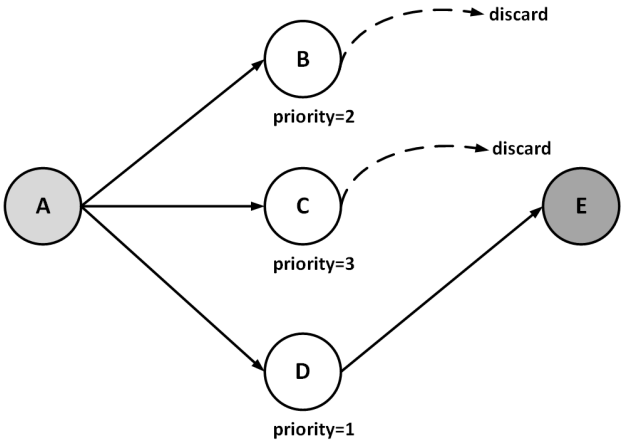

Fig. 5. Example of holding time mechanism

It assumes that nodes A has three candidate forwarders, and the position of B, C and D in priority list are 2, 3 and 1. Upon receiving a data packet by node A, node B, C and D start a timer respectively. According to (9)-(10), we know that node D has a shorter holding time than node B and C. During their holding time, node B and C will simply discard the packet when they overhear the data transmission from node D, which has no additional control packet during this communication process.

V. Algorithm Flow

This paper proposes a Q-learning based routing optimization algorithm for UWSN. The implementation of the algorithm has the following requirements for UWSN:

(1) The node ID in the network is unique;

(2) Nodes are randomly distributed within a specific three-dimensional area and have limited motion;

(3) Nodes can obtain neighbor nodes' relevant information, self-energy and self-position information.

As shown in Fig. 6, the algorithm mainly includes the initialization stage, routing selection stage and maintenance stage, the detailed explanation will be provided later.

A. Initialization Stage

The purpose of this stage is to determine neighbor list and obtain neighbor information through interaction of control packet. The network initialization process is as follows:

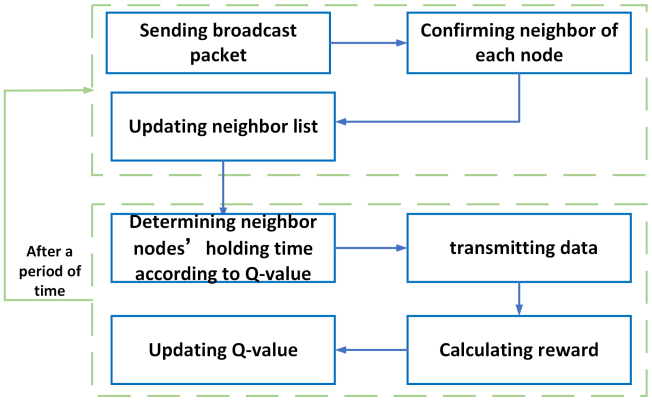

Fig. 6. General views of QLRO algorithm

(1) Each node i broadcasts a HELLO message, which includes the ID of node i and coordinate information (x, y, z), as well as residual energy $E_r(i)$.
(2) If other node j receives the broadcast message sent by node i, they return the ACK confirmation message to i immediately, which contains the ID of node j, residual energy $E_r(j)$, distance $d_{i,j}$ between node i and j, success rate of link $success(i,j)$ and count of data transmission O_j. Fig. 7 illustrated the structure of HELLO packet and ACK packet where header is used to help nodes recognize the type of received packets, and the values in different headers are not identical.
(3) After receiving the ACK message, node i adds the ID number and related information to its neighbor list N(i).

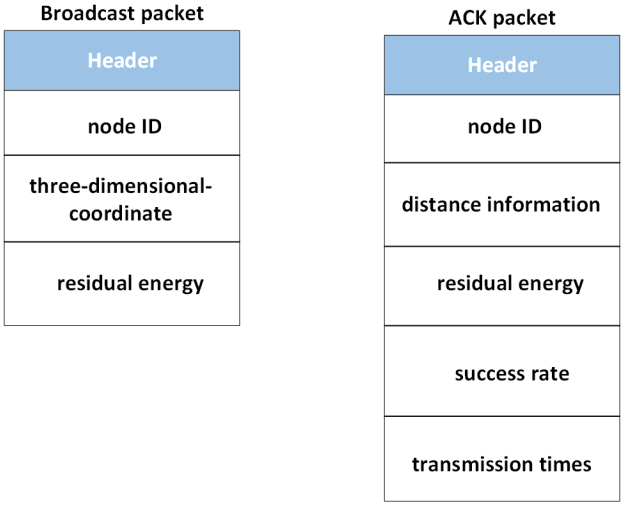

Fig. 7. The structure of interaction packet

B. Routing Selection Stage

When a node requires information transmission, it communicates with its neighbor nodes to update pertinent information. Then, the holding time for each neighbor node is determined based on the current Q-values. After node with a shorter holding time completes data forwarding, other neighbor nodes discard the data packet without repeating forwarding. Finally, update the corresponding Q-value using (8) based on current reward. The above steps will be repeated until the next hop node is surface base station.

Algorithm 1 QLRO algorithm.

initialize Q-value table
while the first node in UWSN dies
  while S is base station s_0
    perceive current state $S = s_t$
    for each node in N(s_t)
      set its holding time according to (9)-(10)
      relayflag=True
      while(timer is not expired) and (relayflag)
        if overhear the packet has been sent
          Relayflag=False
      end while
      if (relayflag)
        transmit the packet
      else
        drop the packet
    end for
    calculate rewards r_t
    a)if s_{t+1} is final state
      $r = r_1$
    b) if s_{t+1} is relay node has no neighbor
      $r = -r_2$
      $flag1 = 1$
    c) if s_{t+1} is relay node and has insufficient energy
      $r = -r_3$
      $flag2 = 1$
    d) if s_{t+1} is relay node
      1) if average residual energy $>$ threshold
        $r = \frac{success(s_t, s_{t+1})}{len/C + d(s_t, s_{t+1})/v_{sound}}$
      2) if the average residual energy $<$ threshold
        $r = \frac{E_r(s_{t+1})}{E_{init}(s_{t+1})} \times \frac{O_{total}}{O_{s_{t+1}}}$
    update Q-values based on current reward
    current state remove to S_{t+1}
  end while
end while

C. Maintenance Stage

When residual energy of a node in network is lower than 10%, the network resends "HELLO" and "ACK" information to re-determine neighbor node. Nodes with low energy are only responsible for sending their own information and do not participate in relaying information for other nodes. Therefore, information of such nodes should be removed from relevant neighbor list.

VI. Algorithm Performance Evaluation

In order to verify the performance of the algorithm in this article, we conducted simulation experiments with MATLAB, and compared with DBR algorithm and QLFR algorithm. Assuming that sensor nodes are evenly distributed in a three-dimensional space of $500m \times 500m \times 500m$. We deploy a base station on the surface. Base station is stationary which energy

TABLE I
SIMULATION PARAMETERS

Parameter	Description	Initial value
v_{sound}	Speed of sound in water	1500m/s
len	Size of data package	random(50B,100B)
v_{min}	Nodes' minimum movement speed	1 m/s
v_{max}	Nodes' maximum movement speed	3m/s
C	Channel capacity	10kbps
$\alpha(f)$	Absorption coefficient	0.002dB/m
P_0	Minimum power required to send	2
P_r	Reception coefficient	0.1

can be continuously supplied by solar energy. Nodes send the collected information to the surface base station through multi-relay nodes. The parameter settings required for simulation are shown in Table I.

In addition, the following four metrics are used to evaluate the performance of the proposed algorithm:

(1) Average end-to-end delay can be expressed as:

$$D_{ave} = \frac{1}{N}\sum_{n=1}^{N} D_n, \tag{11}$$

of which D_n is the delay that it takes to send a packet from the source node and successfully received by the surface station, and N is the total transmission times.

(2) Packet delivery ratio is the proportion of packets successfully received by the base station to the total number of packets sent by the source nodes:

$$PDR = \frac{R_{packets}}{S_{packets}}, \tag{12}$$

where $R_{packets}$ is the total number of packets received by the base station and $S_{packets}$ is the number of packets generated by source nodes.

(3) Total energy consumption refers to the total energy consumption by all nodes in UWSN:

$$totalenergy = \sum_{i=1}^{M}(E_{init}(i) - E_r(i)), \tag{13}$$

where $E_{init}(i)$ is the initial energy of node i, $E_r(i)$ is the residual energy of node i and M is the number of nodes in UWSN.

(4) Network lifetime is the time it takes for the network to run until the first node runs out of energy.

First, we analysis how the performance of QLRO is affected by parameter α. We set the value of α as 0.001, 0.005 and 0.01, respectively.

As shown in Fig. 8, the average end-to-end delay is positively correlated to α. According to (9) and (10), the holding time of a packet will increase with α increasing, therefore the end-to-end delay will increase.

As described in Fig. 9, the packet delivery ratio has negative correlation with α. This is because the decrease of α leads to the reduction of holding time. Therefore, as more nodes participate in packet forwarding, the packet delivery ratio and link transmission reliability will increase.

Fig. 10 represents that the total energy consumption increases while α decreases. When $\alpha = 0.001$, the energy consumption is much more than other two cases. The reason is that, with a smaller α, nodes have shorter holding times, which leads to more redundant packets and more energy consumption.

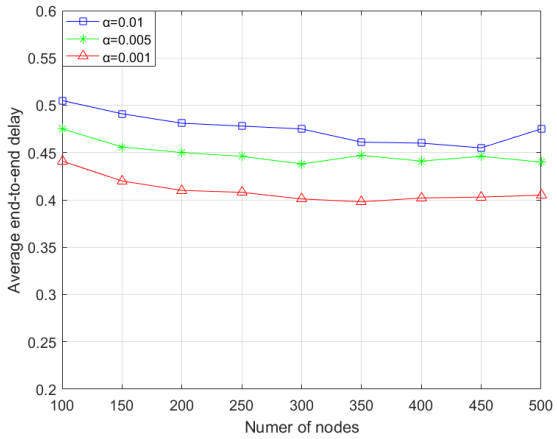

Fig. 8. Average end-to-end delay with varying α

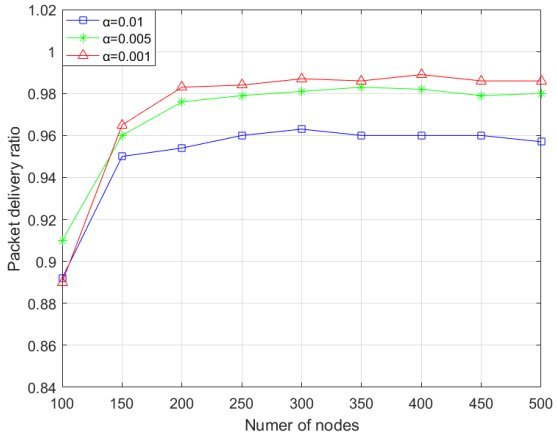

Fig. 9. Packet delivery ratio with varying α

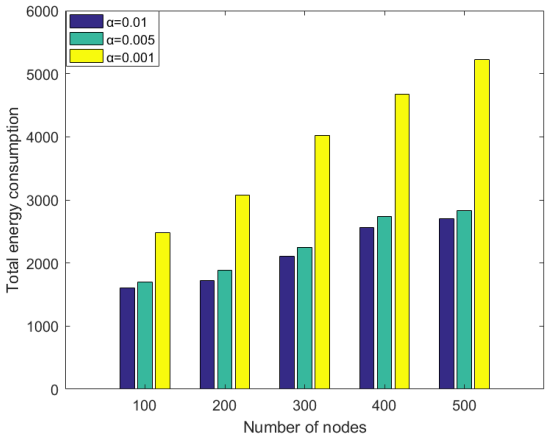

Fig. 10. Total energy consumption with varying α

Now, we compare QLRO algorithm with other routing protocols: DBR and QLFR. From Fig. 11, we can see that end-to-end delay decreases as network density increases and the proposed QLRO has lower average end-to-end delay. For DBR, each node will hold the packet for a certain time, which will increase delay. Although QLFR reduces the overall holding time of packets, it does not fully consider the delay factor in reward function. Moreover, the reward function in QLRO considers transmission delay, which reduces the end-to-end delay effectively.

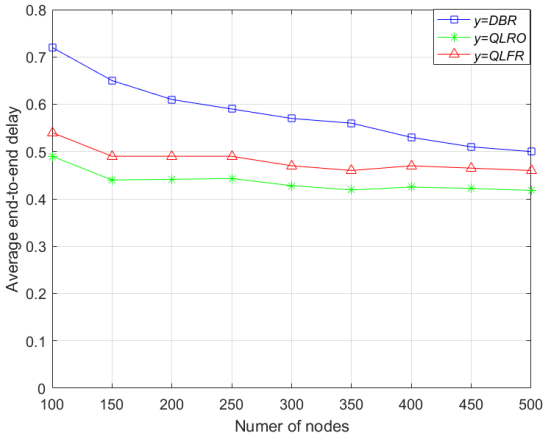

Fig. 11. Comparison of average end-to-end delay

Fig. 12 examines the total energy consumption increases with nodes of UWSN increasing and the total energy consumption of this algorithm is lower than other algorithms. The reason for energy saving is that QLRO can limit the number of nodes participating in packet forwarding and reduce the transmission of redundant packets.

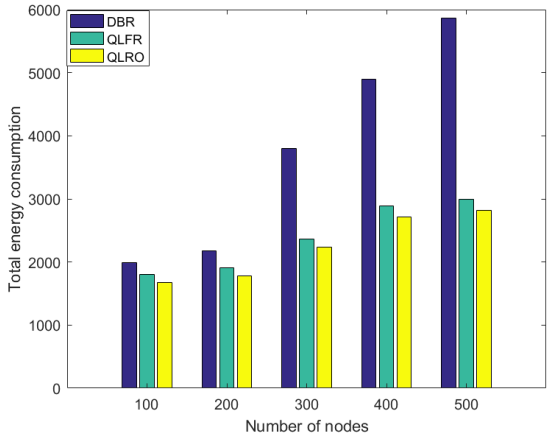

Fig. 12. Comparison of total energy consumption

Fig. 13 examines the network lifetime of DBR, QLFR and QLRO. The network lifetime exhibits a general trend of negative correlation with network density in DBR. This is because DBR can not effectively reduce participation of nodes in packet forwarding. Thus, with the increase of network density, a mass of redundant transmission causes excessive energy consumption, which reduce the network lifetime. While it shows an opposite trend in QLFR and QLRO. QLRO and QLFR restrict participation of nodes in packet forwarding. Furthermore, QLRO combines energy levels and transmissions times in its reward function and selects nodes with higher residual energy levels to transmit data. Therefore, QLRO is more effective in balancing network load and extending network lifetime.

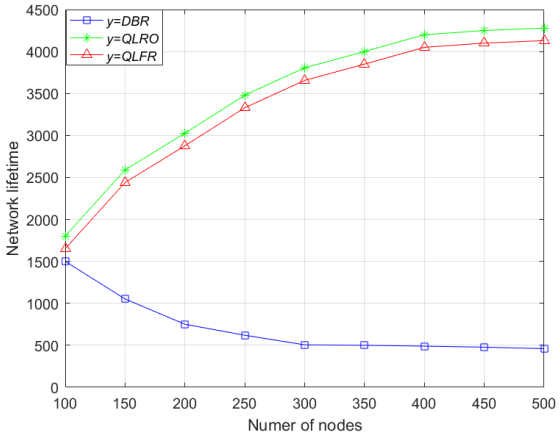

Fig. 13. Comparison of network lifetime

Fig. 14 examines the packet delivery ratio of three methods. The results show that packet delivery ratio of QLRO is lower than DBR and higher than QLFR. In DBR, the enhanced delivery ratio comes at the cost of shortening lifetime. QLRO considers the transmission success rate in its reward function and can better balance the relationship between energy consumption and packet delivery.

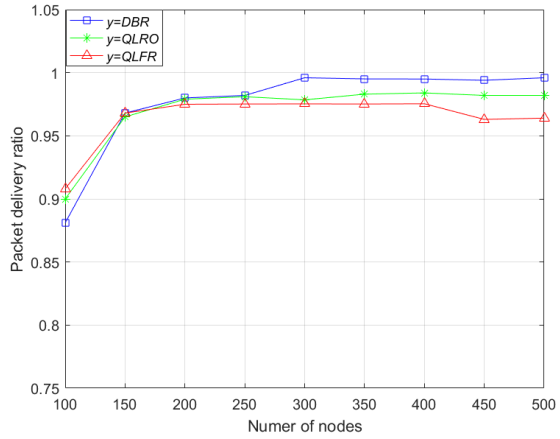

Fig. 14. Comparison of packet delivery ratio

VII. Conclusion

In this paper, we propose a Q-learning based routing optimization algorithm for UWSN. Firstly, we design the delay-based reward and energy-related reward to calculate Q-vale. Then, a novel holding time mechanism based on Q-value is

constructed to avoid duplicate transmission. The performance evaluation reveals that this algorithm can effectively prolong network lifetime, reduce transmission delay and improve package delivery. This algorithm is applicable not only to anchored networks, but also to location-free networks. However, in the latter scenario, underwater precision positioning needs to be addressed to obtain real-time location information of nodes.

REFERENCES

[1] Y. Chen, W. Wang, X. Sun, Y. Tao, Z. Liu and X. Xu. "A cooperative security monitoring mechanism aided by optimal multiple slave cluster heads for UASNs," *China Communications*, vol. 20, no. 5, pp. 148–169, May. 2023.

[2] M. Chaudhary, N. Goyal, A. Benslimane, L. K. Awasthi, A. Alwadain and A. Singh. "Underwater Wireless Sensor Networks: Enabling Technologies for Node Deployment and Data Collection Challenges," *IEEE Internet of Things Journal*, vol. 10, no.4, pp. 3500-3524, Feb. 2023.

[3] A. Hariyale, A. Thawre and B. R. Chandavarkar. "Mitigating unsecured data forwarding related attack of underwater sensor network," in *12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, pp. 1–5, 2021.

[4] Z. Liu, X. Jin, Y. Yang, K. Ma and X. Guan. "Energy-Efficient Guiding-Network-Based Routing for Underwater Wireless Sensor Networks," *IEEE Internet of Things Journal*, vol. 9, no. 21, pp. 21702-21711, Nov. 2022.

[5] K. Tejaskumar, I. R. Dharamdas and B. R. Chandavarkar. "A Mathematical Model for Node Mobility during Water Current and Tsunami in Underwater Sensor Networks," in *11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, pp. 1–6, 2020.

[6] Y. Zhou, H. Yang, Y.H. Hu. "Cross-layer network lifetime maximization in underwater wireless sensor networks," *IEEE Systems Journal*, vol. 14, no. 1, pp. 220–231, March. 2019.

[7] Z. Mohammadi, M. Soleimanpour-Moghadam, M. Askarizadeh. "Increasing the lifetime of underwater acoustic sensor networks: Difference convex approach," *IEEE Systems Journal*, vol. 14, no. 3, pp. 3214–3224, Sept. 2020.

[8] A. Signori, F. Campagnaro, I. Nissen and M. Zorzi. "Channel-Based Trust Model for Security in Underwater Acoustic Networks," *IEEE Internet of Things Journal*, vol. 9, no. 20, pp. 20479-20491, Oct. 2022.

[9] M. Wang, Y. Chen, X. Sun, F. Xiao, X. Xu. "Node Energy Consumption Balanced Multi-Hop Transmission for Underwater Acoustic Sensor Networks Based on Clustering Algorithm," *IEEE Access*, vol. 8, pp. 191231–191241, Oct. 2020.

[10] J. Yang, G. Qiao, Q. Hu, L. Xu, P. Xiao and J. Zhang. "Collision classification MAC protocol for underwater acoustic communication networks using directional antennas," *China Communications*, vol. 19, no. 5, pp. 241–252, May. 2022.

[11] N. Morozs, P. D. Mitchell and R. Diamant. "Scalable Adaptive Networking for the Internet of Underwater Things," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 10023-10037, Oct. 2020.

[12] Z. Yan, X. Yang, L. Sun and J. Wang. "Inter-carrier interference-aware sparse time-varying underwater acoustic channel estimation based on fast reconstruction algorithm," *China Communications*, vol. 18, no. 3, pp. 216–225, March. 2021.

[13] J. Lin, G. Wang, Z. Zheng, R. Ye, R. He and B. Ai. "Modeling and channel estimation for piezo-acoustic backscatter assisted underwater acoustic communications," *China Communications*, vol. 19, no. 11, pp. 297–307, Nov. 2022.

[14] R. Hou, L. He, S. Hu, J. Luo. "Energy-Balanced Unequal Layering Clustering in Underwater Acoustic Sensor Networks," *IEEE Access*, vol. 6, pp. 39685–39691, July. 2018.

[15] A. Bereketli, M. Tümçakır, B. Yeni. "P-AUV: Position aware routing and medium access for ad hoc AUV networks," *Journal Of Network And Computer Applications*, vol. 125, pp. 146–154, 2019.

[16] Z. Jin, Z. Ji, Y. Su. "An Evidence Theory Based Opportunistic Routing Protocol for Underwater Acoustic Sensor Networks," *IEEE Access*, vol. 6, pp. 71038–71047, Nov. 2018.

[17] P. Montague. "Reinforcement Learning: An Introduction," *IEEE Transactions on Neural Networks*, vol. 16, pp. 285–286, 2005.

[18] S. Basagni, V. Di Valerio, P. Gjanci and C. Petrioli. "Harnessing HyDRO: Harvesting-aware Data ROuting for Underwater Wireless Sensor Networks," in *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, Los Angeles, CA, USA, pp. 271–279, 2018.

[19] S. Kim. "A better-performing Q-learning game-theoretic distributed routing for underwater wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 14, no. 1, pp. 1550147718754728, Jan. 2018.

[20] Y. Zhou, T. Cao and W. Xiang. "QLFR: A Q-Learning-Based Localization-Free Routing Protocol for Underwater Sensor Networks," in *IEEE Global Communications Conference (GLOBECOM)*, Waikoloa, HI, USA, pp. 1–6, 2019.

[21] V. Di Valerio, F. L. Presti, C. Petrioli, L. Picari, D. Spaccini and S. Basagni. "CARMA: Channel-Aware Reinforcement Learning-Based Multi-Path Adaptive Routing for Underwater Wireless Sensor Networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 11, pp. 2634–2647, Nov. 2019.

[22] Z. Jin, Q. Zhao and Y. Su. "RCAR: A Reinforcement-Learning-Based Routing Protocol for Congestion-Avoided Underwater Acoustic Sensor Networks," *IEEE Sensors Journal*, vol. 19, no. 22, pp. 10881–10891, Nov. 2019.

[23] S. Wang, Y. Shin. "Efficient Routing Protocol Based on Reinforcement Learning for Magnetic Induction Underwater Sensor Networks," *IEEE Access*, vol. 7, pp. 82027–82037, June. 2019.

[24] X. Li, X. Hu, R. Zhang, L. Yang. "Routing Protocol Design for Underwater Optical Wireless Sensor Networks: A Multiagent Reinforcement Learning Approach," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9805–9818, Oct. 2020.

[25] Y. Lu, R. He, X. Chen, B. Lin, C. Yu. "Energy-Efficient Depth-Based Opportunistic Routing with Q-Learning for Underwater Wireless Sensor Networks," *Sensors (Basel, Switzerland)*, vol. 20, no. 4, pp. 1025, Feb. 2020.

[26] M. Pouyan, A. Mousavi, S.Golzari and A. Hatam. "Improving the performance of Q-learning using simultanouse Q-values updating," in *2014 International Congress on Technology, Communication and Knowledge (ICTCK)*, Mashhad, Iran, pp. 1–6, 2014.

[27] Y. Chen, K. Zheng, X. Fang, L. Wan, X. Xu. "QMCR: A Q-learning-based multi-hop cooperative routing protocol for underwater acoustic sensor networks," *China Communications*, vol. 18, no. 8, pp. 224–236, Aug. 2021.

[28] M. Stojanovic. "On the relationship between capacity and distance in an underwater acoustic communication channel," in *Proceedings of the 1st International Workshop on Underwater Networks*, Los Angeles, CA, USA, pp. 4, 2006.

[29] E. M. Sozer, M. Stojanovic and J. G. Proakis. "Underwater acoustic networks," *IEEE Journal of Oceanic Engineering*, vol. 25, no. 1, pp. 72–83, 2000.

[30] Y. GUO, Y.T. LIU. "Localization for anchor-free underwater sensor networks," *Computers and Electrical Engineering*, vol. 39, no. 6, pp. 1812-1821, 2013.

[31] E. Hyytiä, J. Virtamo. "Random waypoint model in n-dimensional space," *Operations Research Letters*, vol. 33, no. 6, pp. 567-571, 2005.

Jia Gao studied for the M.S. degree in Computer Science and Technology at the School of Information Science and Technology, Qingdao University of Science and Technology from 2019 to 2021. She applied for the Ph.D. degree in 2021, and is currently pursuing her Ph.D. degree at the School of Information Science and Technology, Qingdao University of Science and Technology. Her research interests include underwater wireless communications, underwater network optimization.

Jingjing Wang (Member,IEEE) received the B.S. degree in industrial automation from Shandong University, Jinan, China, in 1997, the M.Sc. Degree in control theory and control engineering Qingdao University of Science and Technology, Qingdao, China, in 2002, and the Ph.D. degree in computer application technology, Ocean University of China, Qingdao, China, in 2012. She is a Professor with the School of Information Science and Technology, Qingdao University of Science and Technology. Her research interests include underwater wireless sensor network, acoustic communications, ultra-wideband radio systems, and MIMO wireless communications.

Jianlei Gu received his bachelor's degree in Communication Engineering at the School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications in 2021. He applied for the M.S. degree in 2021, and is currently pursuing his M.S. degree at the School of Information Science and Technology, Qingdao University of Science and Technology. His research interests include underwater wireless communications, hydroacoustic communication equalization.

Wei Shi is currently an Associate Professor with the School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, China. He received the B.S. degree from Ludong University, Yantai, China, in 2009, and the master's degree in signal and information processing and the Ph.D. degree in computer application technology from the Ocean University of China, Qingdao, China, in 2011 and 2014, respectively. His research interests include underwater wireless sensor network and underwater acoustic communication.