International Core Journal of Engineering Volume 7 Issue 3, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0002

Dynamic Routing Optimization Strategy based on Reinforcement Learning

Dongling Zou, Yangxia Lu

College of Information Engineering, Shanghai Maritime University, Shanghai, China.

Abstract

A reinforcement learning-based dynamic routing optimization method is proposed for traditional routing algorithms that cannot cope well with dynamic network topology changes and congestion and time delay problems in the network. Using a self-learning process, routing is performed in dynamic complex networks using reinforcement learning, and a Q-network-based algorithm is designed to select the next-hop route for data transmission tasks by obtaining information from neighboring wireless sensor nodes, providing feedback on the different choices, In the feedback reward function, multiple factors such as the length of the data transmission path, the delay and the size of the node queue are designed to find the path after the dynamic routing optimization in the network. Simulation results show that the algorithm can adjust the routing strategy adaptively according to the dynamic changing environment of the network, reduce data congestion and transmission delay.

Keywords

Reinforcement Learning; Dynamic Routing; Q-networks; Path Selection; Congestion.

1. Introduction

Wireless communication routing protocol is one of the standards of WLAN. It is a multi-hop self-organizing distributed network formed by sensor nodes through wireless communication, which is responsible for forwarding data packets from source nodes to destination nodes through the intermediate nodes of the network. Routing protocol needs to solve the problem of data transmission. It has the functions of sensing network topology changes, establishing, updating and maintaining routing information in time to ensure efficient, stable and correct data forwarding of wireless sensor nodes. In recent years, the research of machine learning in routing protocols has gradually become a hot topic and attracted much attention. Compared with traditional algorithms, machine learning has some advantages in solving congestion, delay, path selection, packet loss rate, traffic control and signaling overhead.

Papers [1]-[5] use a machine learning approach in static network topologies to learn routing by different optimization objectives, which makes up for some shortcomings of traditional algorithms. However, it does not consider the complex dynamic network topology, does not make pre-assumptions about the dynamics and unpredictability of network topologies. It is only applicable to the fixed static network topology under set conditions, and is not suitable for routing selection when the network changes rapidly.

In the dynamic network topology, the dynamic mobility of nodes and the dynamic length change of links need to be considered. In the related research, this paper [6] adopted the reinforcement learning method based on trial and error, combined with the mobility of nodes, and selected the best scheme from all neighbors to send data packets to the target node, so as to improve the delivery rate of packets and reduce the transmission delay. In the paper [7], a routing scenario based on reinforcement learning

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0002

and dynamic change of node position was proposed to reduce the average delivery time in the case of high traffic. The paper [8] proposes Q-learning based adaptive routing, which uses reinforcement learning to detect the ability of node movement levels at different times, enabling nodes to update routing information, reducing latency while reducing packet loss. In this paper [9], deep learning method is introduced to control the movement and data forwarding of mobile sensor nodes, so as to meet the needs of network data transmission and monitoring. The above papers use machine learning to simulate the dynamic change of network topology by changing the location of nodes, consider the feasibility in dynamic routing. However, the dynamic change of link length is not taken into account.

Based on the above analysis, this paper uses reinforcement learning method to solve the routing protocol problem under dynamic network topology when the link length changes dynamically. By designing the Q-learning algorithm model, the delay, congestion and path length are taken as the optimization objectives, and the optimal combination is constructed to obtain the action strategy. Then, it interacts with the environmental information, perceives the system routing state and environmental changes, and records the feedback information generated after the action, and finally converges to the optimal action through iteration. Compared with the traditional routing optimization algorithm, the intelligent dynamic routing optimization algorithm Q-routing proposed in this paper uses the intelligent algorithm model to train, which endows the wireless routing sensor nodes in the network with self-learning ability. According to the dynamic changes of input and network environment, the optimized routing decision can be obtained by fast reasoning, the intelligent routing method can better adapt to the dynamic network application scenarios, complete more and more complex data transmission tasks, achieve the routing optimization goal, effectively avoid network congestion and reduce data transmission delay. At the same time, the random dynamic change of link length in network topology greatly enriches the unpredictability of network structure, and makes the intelligent routing method have better scalability and diversity in the process of deploying network topology. Finally, a simulation system based on dynamic network topology is built to verify the feasibility and effectiveness of the proposed method.

2. System model

2.1 Dynamic network topology model

Topology including static and dynamic two kinds of structure, heterogeneous static topology is often used in research, but in practice, topology is not always static, and its traffic patterns and link changes are time-varying. Therefore, it is necessary to consider the network structure with dynamic changes. In the dynamic and complex network structure, the failure of a node or the failure of the connection link between nodes may occur in the network [10]. Therefore, routers need to acquire changes in the network topology and alter the path of distributed packets to quickly adapt to the new network topology.

In this paper, BA scale-free stochastic dynamic network topology is adopted. We suppose the number of nodes already in the network is s. In the process of dynamic network change, a node with index subscript m is added, and the probability that the new node m is connected to node v already in the network is p, defined as

$$p(v) = \frac{k_v}{k_1 + k_2 + \cdots k_s} \tag{1}$$

In the formula, k_1, k_2 and k_s respectively represent the number of adjacent nodes that can be connected when a node joins the network. The probability of new node m and node v can be shown as the following Fig. 1.

Fig. 1. Probability distribution

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0002

After calculating the probability of connection, random numbers between 0 and 1 are randomly generated by the network to act on the probability interval, and the final decision is made on which node in S to connect. This can ensure that the connection between the introduced new node and the existing node in the network is random. After T time steps, the total number of network nodes is S+T.

In the dynamic network topology, node between the link length changes there are three ways: Random deletion of links, random retention of links, and random change of partial link length. Rules for randomly changing links are set as formula (2), where X_{mv} represents the initial link length between adjacent nodes m and v, θ and X represent random parameters, $\theta \in (0, \pi)$.

$$\hat{X}_{mv} = \begin{cases} X_{mv} * [1 + \cos\theta] & change \\ X_{mv} & preserve \\ None & delete \end{cases} \quad (2)$$

Where $\hat{X}_{mv}$ represents the length updated by X_{mv}; for the convenience of calculation, the random length of the initial network link $X \in (2,7)$.

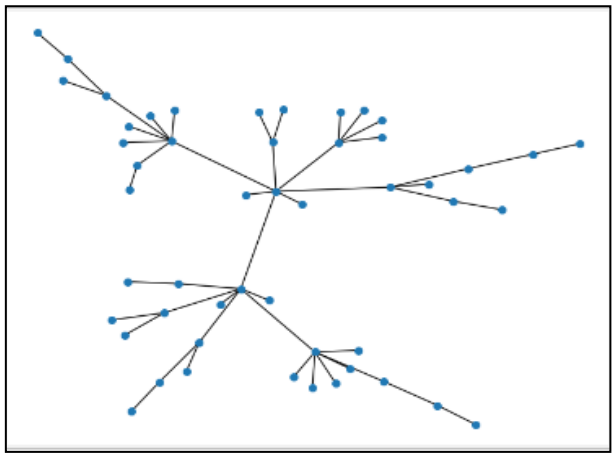

Fig. 2. Network topology

When the number of network nodes is S=39 and the number of new nodes is 1, the network topology is shown in Fig. 2 in a radial form. where the node connection method and link change rules are the same as described in equation (1) and (2).

2.2 Q -learning model

Q-learning is a reinforcement learning optimization algorithm [11]. The agent is constructed. It is assumed that the agent perceives the environment at t time, selects the action a_t under the state s_t as the state-action pair (s_t, a_t), and then the agent gets the reward r_{t+1}, and the environment enters the next state s_{t+1}. Through the interaction with the environment, the agent continuously learns the action to maximize the future reward without knowing what state the environment will change to or which behavior will produce the highest reward, and obtains the optimal strategy.

The defined current strategy π(s) is the action a_t at time t. The value of Q obtained by state s_t under action a_t at time t is the maximum value of the Q value obtained by all possible actions of state s_t under the current policy. The current policy at this time is the optimal policy $\pi'(s)$. It can be expressed as:

$$\pi'(s) = argmax Q_\pi(s_t, a_t) \quad (3)$$

Where Q_π represents the quality assessment of the actions under the current strategy.

The core idea of Q-learning is to approach Q function iteratively through Bellman equation, and the updated formula of Q function is shown in (4):

$$Q_\pi(S_t, A_t) \leftarrow Q_\pi(S_t, A_t) + \alpha\,[R_t + \max_a Q_\pi(S_{t+1}, a) - Q_\pi(S_t, A_t)] \quad (4)$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0002

In the formula, $Q_\pi(S_t, A_t)$ denotes the quality of action A executed under state S. The Q value is calculated based on value iterations. The asymptotic method is used to make a small step towards the target, and finally gradually converge the Q function. The α in the algorithm is the learning rate and is used to indicate the extent to which the newly obtained Q value changes the previous Q value. The $\max_a Q_\pi(S_{t+1}, a)$ used to update $Q_\pi(S_t, A_t)$ represents the largest of all actions in the next state S_{t+1} to get the Q value by looking up the current Q table. The Q function estimates the merit of the choice of action A_t under state S_t by the value of $Q_\pi(S_t, A_t)$, which is an approximation. However, with the increasing number of iterations, the approximation becomes more and more accurate, when the iteration reaches a certain number, the Q function converges to the real Q value.

3. Dynamic routing strategy based on Q-Learning

This section fully combines the dynamic network topology model and reinforcement learning Q-learning model, and uses the Q-learning-based model to study the routing strategy suitable for dynamic network topology, as shown in Fig. 3. The network topology information and network environment information in the event are constantly changing dynamically. The current wireless sensor node processes the relevant information received by neighbor nodes in real time and forwards the data to the next routing node to get feedback and obtain reward. The Q value is calculated based on feedback to judge the quality of the next step of routing decision, update the routing table in time, dynamically adjust the appropriate routing and forwarding strategy in accordance with the overall network environment.

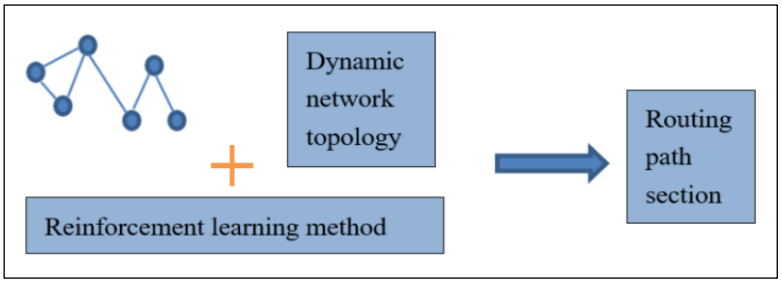

Fig. 3. Q-routing dynamic routing structure

3.1 Routing and forwarding strategy

The routing process of wireless sensor nodes in the dynamic network is the process of selecting the next-hop routing for data forwarding, which can be regarded as the routing problem of Markov decision process following the dynamic programming principle. Throughout the network, the routing problem is given a dynamic Q table to restore and update. Nodes act as agents to reduce congestion, transmission path length and delay, and at the same time, they will also express their views of the network and update the Q table. Obtaining data information of neighboring nodes to select actions, get feedback, and then learn to adjust routing strategy. The mathematical expression is as follows:

$$Q_{cj\ markov}^{t+1} = E_{j\ \epsilon\ neighbor\ y}[j] \tag{5}$$

In the formula, the current node c obtains the information of neighboring nodes to make an action, and selects the next-hop routing target node j to carry out data forwarding and obtain information feedback. $Q_{cj\ markov}^{t+1}$ represents the update of Q table after the current node c receives the feedback from the target node j at the time of t+1, and E represents the feedback information of the target node j.

The random change of link length in network topology follows the updating change of Q table. The faster the update of Q table is, the faster the random change of link will be. The dynamic routing network topology used in this paper is shown in Fig. 2 before the event does not start, and the link change follows the formula of Section 2.1 (2).

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0002

The entire network selects the route in the unit of time slot. Within each time interval, the next hop router is selected, and the packet is sent out and stored in the cache queue of the next hop router. Whether the packets that the current node needs to transmit exceeds the remaining buffer size of the next hop router as the basis for network congestion.

In the process of continuously forwarding data, the node completes the Q-learning algorithm training through self-learning. In view of the dynamic network topology and network environment, the agent changes the routing forwarding strategy in real time, updates the Q table, and avoids congestion and so on. As the learning process goes on, the performance of the entire network system improves from the learning process.

3.2 Model training

The routing problem needs to be defined as follows:

1) Agent: Its purpose is to maximize the cumulative reward while reaching the destination node, starting from the start node and considering the best performance path forwarding.

2) State: The state in the model can be regarded as the position information of the current node c and the destination node d, which can be represented by set S, and g represents the number of forwarding nodes. In network events, the number of forwarding nodes represents the number of states.

$$S = \{(c_1, d) \dots \dots (c_g, d)\} \tag{6}$$

3) Observation: represents the observed network environment, including whether the current node has adjacent nodes y and the queue size q of adjacent nodes in the changing topology, which can be represented by set O.

$$O = \{(y, Q_1,) \dots \dots (y, Q_g)\} \tag{7}$$

4) Action: Select the next-hop router. It can be represented by the set A.

$$A = \{a_1, a_2, \dots \dots a_{g-1}\} \tag{8}$$

5) Reward: It can be set as a function of integrating different indicators according to different network optimization needs, which can be represented by set R.

a. If the action is none (the current node has no neighbors or the given neighbors have no space in the queue), $reward_1 = 0$;

b. If the packet is transmitted to the destination node, the reward function r_0 is given;

c. If there is no available path to the destination node (i.e., there is no available path between the starting node and the destination node), then $reward_2 = -40$, the number 40 represents the total number of nodes in the network structure;

d. In other cases, the queue size of the target node b_j; after selecting the action, the transmission delay T_{cj} between leaving the current node c and reaching the target node j; and the link length X_{jd} between the target node j and the destination node d are taken into account according to the network optimization requirements. The expressions are as follows:

$$X_{jd} = ||node_j - node_d|| \tag{9}$$

$$reward_3 = -(b_j + w_1 T_{cj} + X_{jd}) \tag{10}$$

Where, w_1 represents the adjustment parameter.

Get the formula:

$$R(s, a, s') = \begin{cases} r_0, & arrive\ destination \\ r_1, & action\ is\ none \\ r_2, & no\ \ path \\ r_3, & arrive\ target\ node \end{cases} \tag{11}$$

6) Training Update:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0002

The current node c transmits data to node j at time t, and node j feeds back information to node c at time t+1. $R_{t+1}^{c}(c,j)$ represents the feedback reward that node c selects the next hop target node j to send information. With feedback reward, node c can update relevant information of the dynamic routing table with Equation (12):

$$Q_{t+1}(c,j) \leftarrow Q_{t+1}(c,j) + \alpha[R_{t+1}^{c}(c,j) + \gamma \max_{a} Q_t(c',j') - Q_{t+1}(c,j)] \quad (12)$$

Where: c' and j' respectively represent the current node and target node in the next time slot.

Through the feedback information of node j, node c can learn the good and bad quality of data transmission to target node j. By continuously obtaining feedback information, nodes get updated network information, so as to adjust the routing strategy in dynamic complex networks. α represents the rate of updating the routing table, γ represents the reward discount factor.

3.3 Obtain the optimal combination path

After training and updating according to the above (12), the routing algorithm is trained to be good enough to approach the convergence state, the agent learns the best path of the data transmission task from the source node to the destination node, and finds the optimal routing optimization strategy to maximize the cumulative reward. The mathematical expression of the optimal path is as follows:

$$P^* = \{c_1 c_2 \dots c_g | k = 1 \dots g, max \Sigma_{k=1}^{g} r_k\} \quad (13)$$

In the formula, P^* represents the optimal combination path, g represents the number of nodes used, and r_k represents the reward for selecting k node.

3.4 Evaluation index

After the end of the whole dynamic Q network process, the average delivery time of N packets in the network and various measures of network congestion are calculated to verify and evaluate whether a better routing optimization can be made in the dynamic network and whether it has good performance. That is to reduce the number of packet loss, shorten the path length of data transmission tasks, reduce the average transmission time T, alleviate congestion. The formulas for average transmission delay and congestion metrics are as follows:

Average data transmission delay calculation formula:

$$T = \frac{1}{N} \sum_{n \in N} t_n \quad (14)$$

Where, t_n represents the total transmission delay of data packets, where N represents the number of loads.

Mathematical expression of congestion measure W:

$$W \in (H, E)$$

Where H represents the proportion of empty nodes and E represents the average queue size of nodes.

The formula for calculating average queue size of nodes is as follows:

$$E = \frac{\mathrm{B}}{\mathrm{g}} \quad (15)$$

Where: B represents the total queue size of selected nodes, and g represents the selected nodes.

4. Simulation and analysis

In this paper, a routing path selection method based on RL (reinforcement learning) in dynamic complex networks is proposed. The Q-learning algorithm, which explore the probability of using random action selection epsilon is set to 0.9, and the python simulation platform is used for simulation. The proposed algorithm and the shortest path algorithm based on traditional routing OSPF, are compared and analyzed from multiple performance to verify the effectiveness of the proposed algorithm. Before the iteration, network load(packet) has been generated in the network. After the beginning of the iteration, the network topology links change randomly and dynamically, and the starting node and destination node of the data transmission task are randomly selected. The simulation

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0002

calls nodes to determine the combined path for each packet according to the routing algorithm. When the set packets are generated and forwarded, the whole learning iteration process ends. The following figures show the comparison of the simulation results.

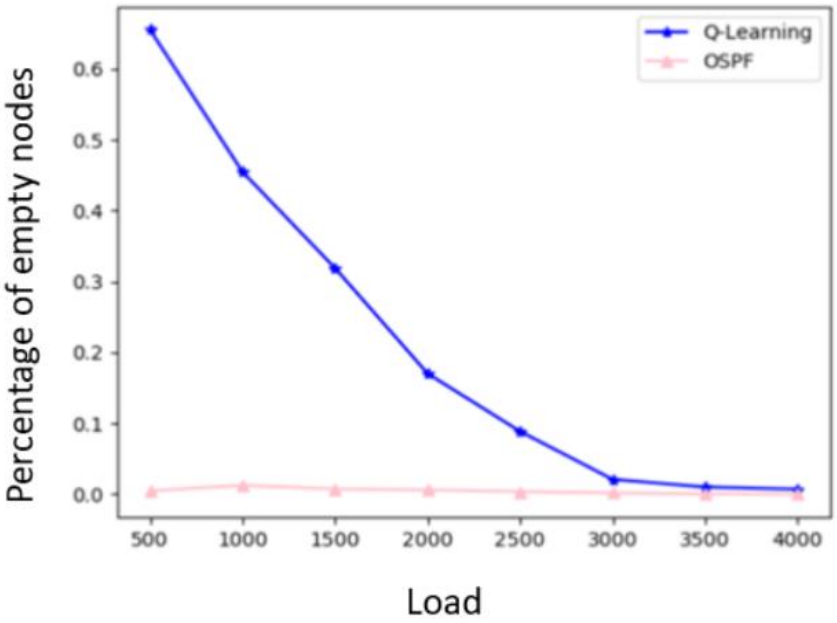

Fig. 4. Percentage of nodes empty and load

The relationship between the proportion of empty nodes and the number of network load is shown in Fig. 4. The routing selection algorithm based on Q-learning is relatively less used on nodes and more sensitive to the network environment. When selecting the next hop node, the remaining path length is fully considered by information feedback. Compared with the traditional shortest path routing algorithm, it has obvious advantages and can reduce the transmission path. When the load is less than 3000, the algorithm has good adaptability, and the use of nodes is proportional to the load. However, the traditional OSPF algorithm has a high utilization rate for the nodes in the network, almost 100%.

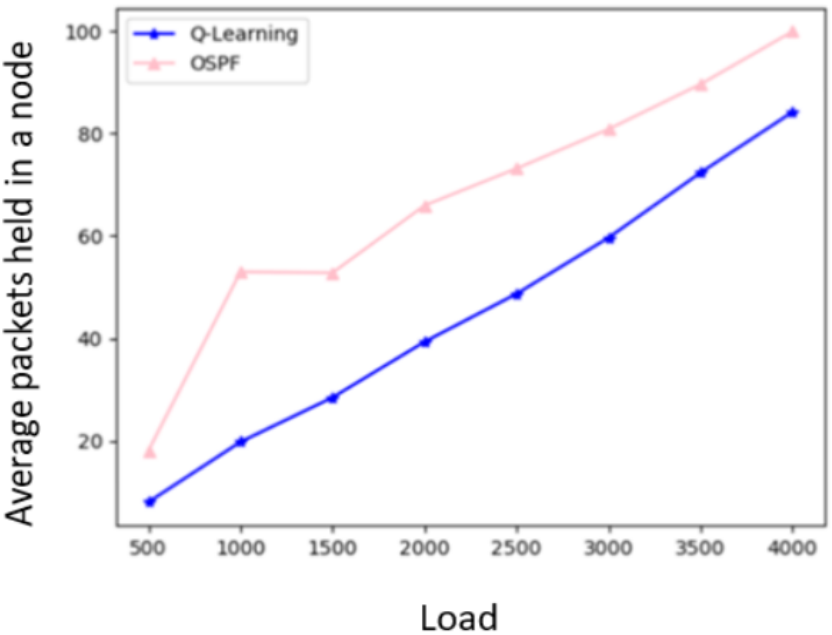

Fig. 5. The average number of packets held in a node and the load

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0002

The relationship between the average queue size of nodes in the network and the load is shown in Fig. 5, and the Q-learning method has relatively good overall performance in dynamic routing. load at the beginning of 500, the gap is not big, using fewer nodes. As the progress of simulation, the queue space of OSPF algorithm increases significantly with the increase of load. Finally, the proposed algorithm is consistently about 30 packets less than the traditional algorithm. It is shown that the proposed method has good data transmission ability and can reduce congestion.

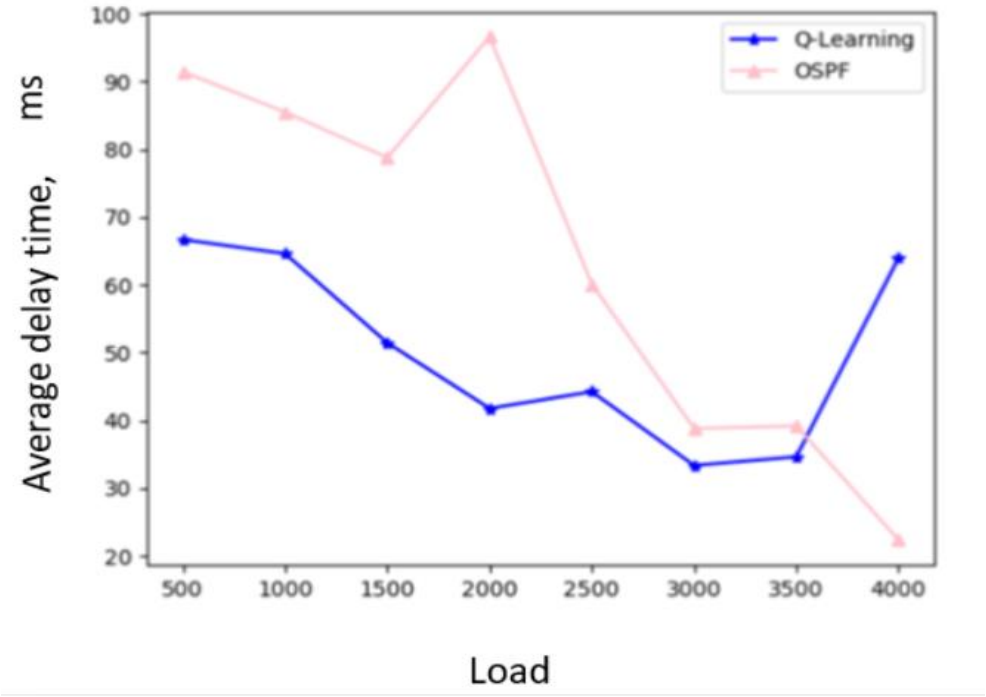

Fig. 6. Relationship between average transmission delay and load

The relationship between average packet delivery time and network load is shown in Fig. 6, compared with the traditional routing algorithm, the Q-learning-based routing algorithm has lower time delay at the beginning, which indicates that the proposed algorithm has better adaptability to dynamic network environment and can quickly make optimal path decisions. Update routing information in time, reduce delay caused by path congestion; When the load increases to 1500, the transmission delay of the traditional algorithm increases, while the proposed algorithm maintains a decline and has good stability. When the network load is less than 3500, the average delay of the proposed algorithm is less than that of the traditional algorithm, with obvious difference and good performance. However, when the load is 4000, the proposed algorithm will show some limitations.

5. Conclusion

This paper combines reinforcement learning model Q-learning and dynamic network topology model, and takes wireless sensor nodes in dynamic network topology as the research object. The nodes obtain the data information of adjacent nodes in the dynamic network, and select the next transmission path of the packet through the reinforcement learning algorithm, so as to obtain the feedback reward. The setting of the objective function in the feedback reward integrates multiple optimization indexes. The nodes use the feedback information to update the dynamic routing strategy to complete the learning process and find the optimal combination path for the data transmission task. The simulation results show that, compared with the traditional algorithm, Q-learning is superior to the traditional shortest path algorithm in the performance of the dynamic topology structure of complex networks and the dynamic updating of the routing table, which effectively reduces the delay and alleviates the congestion. At the same time, Q-learning has strong topology adaptive ability and fast topology

ISSN: 2414-1895 DOI: 10.6919/ICJE.202103_7(3).0002

change perception sensitivity. However, when the network load is too large, it will show some limitations. Such as the routing strategy in the network is unable to be fine-tuned, and the adaptability to network load changes is not strong. It may not be able to optimize multiple target requirements at the same time. Next, we will explore the deep learning method to more adapt to the changes in the dynamic network topology structure, and use the function approximation ability of deep learning to solve the large-scale state and action space storage problems, so as to avoid the problem of maintaining the storage resources caused by large-scale Q tables and reduce the time cost of table lookup, in order to explore better routing path decisions.

References

[1] N. Kato, Z. Fadlullah, B. Mao, F. Tang, Q. Akashi, T. Inoue, K. Mizutani, "The deep learning vision for heterogeneous network traffic control: Proposal, challenges, and future perspective," IEEE Wireless Communication, vol. 24, no. 3, pp. 146–153, Jun. 2017.

[2] Kiani F, Amiri E, Zamani M ,et al. Efficient intelligent energy routing protocol in wireless sensor networks. International Journal of Distributed Sensor Networks .2015.

[3] F. Tang et al., ''On removing routing protocol from future wireless networks: A real-time deep learning approach for intelligent traffic control,'' IEEE Wireless Communication., vol. 25, no. 1, pp. 154–160, Feb. 2018.

[4] Zubair M, Baomin M et al., State-of-the-Art deep learning: Evolving machine intelligent toward tomorrow's intelligent network traffic control systems,2017, IEEE.

[5] Boriboon D et al., Dynamically Packet routing for QoS assurances on internet networks. 2018, IEEE.

[6] Ghaffari A. Real-time Routing algorithm for Mobile AD Hoc Networks Using Reinforcement Learning and heuristic algorithms[J]. Wireless Networks,2017,23(03):703-714

[7] S.K. Routray, Sharmila K.P. Routing in dynamically changing node location scenarios: A reinforcement learning approach. 2018, IEEE.

[8] Serhani A, Naja N, Jamali A. QLAR: A Q-learning based adaptive routing for Manets. 2016, IEEE.

[9] Oda T, Obukata R, Ikeda M, et al., Design and Implementation of a Simulation System Based on Deep Q-Network for Mobile Actor Node Control in Wireless Sensor and Actor Network. 2017:195~200.

[10] Antonio Mira Lopez, Douglas R., Simulated Annealing Based Hierarchical Q-Routing: A Dynamic Routing Protocol. 2011, IEEE.

[11] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction. Cambridge, MA, USA: MIT Press, 2018.