Joint Federated Learning and Reinforcement Learning for Maritime Ad Hoc Networks: An Integration of Personalized Collaborative Route Planning

Chengzhuo Han[1](✉), Tingting Yang[2,3], and Huapeng Cao[3]

[1] School of Cyber Science and Engineering, Southeast University, Nanjing, China
hcz_dmu@163.com
[2] Peng Cheng Laboratory, Shenzhen, China
[3] National Institute of Defense Technology Innovation, Academy of Military Science China, Beijing, China

Abstract. Maritime Ad hoc networks are a type of decentralised wireless network with rapid networking and multi-hop routing, which are independent of fixed base stations. Recently, Ad hoc networks have started to play an increasingly important role in military command, emergency rescue, disaster relief, temporary meetings, and other occasions. However, as the network topology changes rapidly and the node energy and network bandwidth are limited, discovering and maintaining reliable transmission paths have become a highly topical challenge. In order to solve the problem that distributed routing planning of large-scale Ad hoc networks cannot adapt dynamic changes in network topology, and considering the differences of network nodes, this paper proposes federated reinforcement learning to improve the efficiency of distributed routing planning through the joint learning of similar nodes. Different network nodes have different routing policies, but the routing tables of neighboring nodes are very similar. Therefore, our federated reinforcement algorithm learns nodes with similar routing policies. In this study, a communication system simulation software is specially designed to evaluate the performance of the proposed algorithm.

Keywords: Federated reinforcement learning · Routing planning · Maritime ad hoc network

1 Introduction

Ad Hoc networks are a distinct type of wireless communication network. And Ad Hoc networks has a certain flexibility in the networking process and a reasonably strong ability to adapt to the environment relatively fast. Within a limited area,

Supported by organization 2020YFB1806800.

L. Wang et al. (Eds.): WASA 2022, LNCS 13472, pp. 279–290, 2022.
https://doi.org/10.1007/978-3-031-19214-2_23

more mobile conditions can be provided to improve the working environment for the operation of mobile communication equipment and meet specific work needs. Ad Hoc networks can also be widely used to provide wireless network support in disaster rescue, remote area development, national defence [11], campus teaching [13] and maritime communications. In wireless maritime Ad Hoc networks, network communication depends on the cooperation between vessels and information forwarding between vessels [2]. As vessels move, the network topology changes dynamically. In wireless self-organising networks, all vessels have equal status and virtually the same complexity. Two vessels that are far away and cannot communicate directly can forward control and data messages via multi-hop relay to complete the communication process. Wireless maritime Ad Hoc networks have enormous potential, which can be better applied in various communication fields.

The deployment of multi-hop relay and forwarding has broad future application prospects, particularly in deep-sea areas where there are few users as it can save deployment costs and makes data transmission between users more flexible. However, many problems related to the reliability of multi-hop relay transmission still need to be solved to ensure the reliability of service transmission, especially how to avoid packet congestion in the network. To this end, some recent works have proposed various solutions [5].

Under the new situation, Ad Hoc network communication can be regarded as a layered control network system composed of multiple agents, which adopts edge computing and relies on the distributed parallel mode among intelligent groups to share information and make collaborative decisions, and finally completes the communication task [8]. At the same time, edge computing reduces the network communication load and improves the system operation efficiency through independent decision-making, key information sharing and task collaboration. In the process of multi-agent execution, cooperative and efficient routing is crucial to improve network performance. This problem is called multi-agent communication planning. Designed to generate good communication routes that guide packets from the source node to the specified destination node.

Recently, many scholars have solved large-scale problems by assigning global control to local agents, which is a significant improvement over centralized reinforcement learning [3]. Unfortunately, in the case of limited communication, each agent is only partially observable of the environment, so it is easy to fall into local optimality. However, for large collaborative communication problems, the centralized RL approach is usually not feasible because: 1) Collecting all the maritime observations in the network to form a global state, which in practice causes high latency; 2) The joint action space of each agent grows exponentially with the increase of the number of agents. Therefore, it is more effective and reasonable to make the large-scale cooperative communication as a cooperative multi-agent decision-making system, that is, each agent controls by local observation.

Distributed wireless maritime Ad Hoc networks use distributed scheduling [9], where nodes share local observations to avoid congestion during message transmission. In distributed networks, nodes only need to maintain and forward the information of neighbour nodes to complete resource scheduling, therefore reducing frequent signalling forwarding between nodes, and greatly reducing overheads compared with centralised networks [6]. Therefore, the in-depth study of distributed wireless multi-hop maritime Ad Hoc networks is of great significance to the development and future application of wireless communication networks.

We treated each node of the maritime Ad Hoc network as an agent and transformed the routing planning problem into a multi-agent communication problem. This paper combines reinforcement and federated learning and proposes that the resulting combined federated reinforcement learning should be combined to solve the above issues. In reinforcement learning to learning as a testing evaluation process, the agent chooses an environment action and the environment, after accepting the action state change, simultaneously produces a strengthening feedback signal (award or punish) to the agent.

Federated learning stores the data of each node locally so the federated system can establish a virtual common model without violating data privacy laws and regulations by exchanging encrypted parameters [7,12]. In this paper, the actions selected in federated reinforcement learning (FRL) not only affect the current node reinforcement value, but also affect the neighbouring states and final reinforcement value. This virtual model is in effect a combined optimal model; however, when creating virtual models, the data itself does not move, nor does it compromise privacy or affect data compliance. In this way, constructed models achieve adjacent region goals in their respective regions. The main contributions of this paper are as follows.

1. **Modeling and Formulation:** We formulate the distributed joint routing problem under maritime and network as markov decision process. For distributed decision making, we aim to reduce the total cost of communication computation while considering the impact of other agents' decision results on the current agent. In addition, we hope that the algorithm can take into account the similarity and difference between nodes.
2. **Algorithm Design:** Joint reinforcement learning proposed by us can improve the efficiency of distributed routing planning through joint learning of similar nodes. Considering the differences of network nodes, it solves the problem that distributed routing planning of large-scale Ad hoc networks cannot adapt to the dynamic changes of network topology.
3. **Experimental Verification and Evaluation:** We performed extensive simulations to evaluate the FRL algorithm. The simulation results not only verify the theoretical tradeoff of FRL, but also show that the FRL algorithm can effectively reduce the total cost of the system and improve the level of algorithm personalization.

The remainder of this paper is organized as follows. A typical mobile maritime Ad Hoc network model is given in Sect. 2 and problem formulation is presented in Sect. 2.4. FRL algorithm is proposed, as well as its advantages in processing heterogeneous data are demonstrated in Section in Sect. 3. In Sect. 4, Simulation results of packet routing planning demonstrate the superiority of the proposed method. We conclude this paper with future work in Sect. 5.

2 Problem Formulation

Maritime Ad Hoc networks receive signals wirelessly. Information can be forwarded to other nodes beyond the wireless transmission range of its own node, that is, any network topology can be formed through wireless connection. It is also a self-organising, infrastructure-free wireless network.

2.1 The Maritime Ad Hoc Network Model

A typical mobile maritime Ad Hoc network model is shown in Fig. 1. In this model, every node in the network is mobile, there is no fixed infrastructure, and the status of nodes is equal. Each node (mobile terminal) is responsible for forwarding packets, finding routes, and maintaining paths. A node faces both a user and a device. Due to the wireless coverage of nodes, fixed object blocking and other reasons, communication between nodes in maritime Ad Hoc networks is generally multi-hop. As shown in Fig. 1, nodes A and I cannot communicate directly, but can communicate through the path A-B-D-F-I.

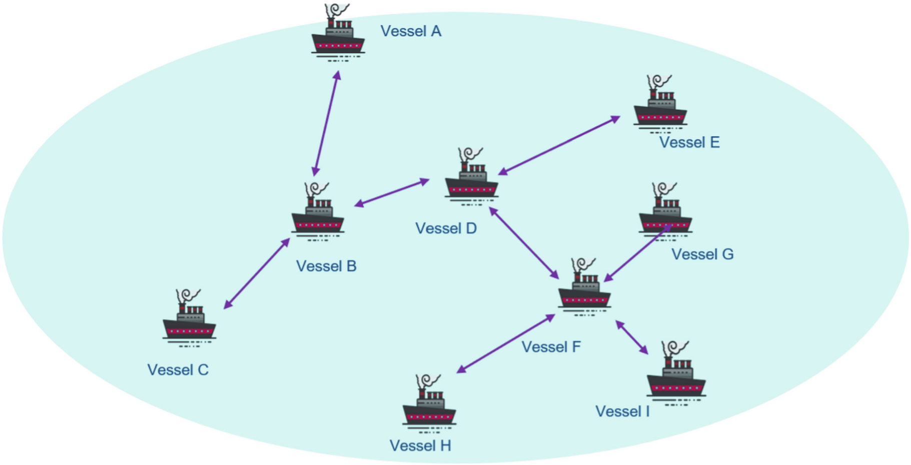

Fig. 1. An illustration of Ad Hoc network

2.2 Data Packet

Network data is transmitted in packets, and each packet has a sending node, destination node, and a current node [10]. We use the arrival time and arrival rate of packets to evaluate routing decisions. If the packet arrives at the destination node before the specified time or exceeds the specified time, the current packet will be deleted, and new packets will be injected into the network. The data packets $\mathcal{P} = \{P_1, \cdots, P_n\}$ can be transmitted on nodes $\mathcal{J} = \{J_1, \cdots, J_m\}$. The packet has parameters $i, j, c \in \{1, \cdots, m\}$, where i represents the sending node, j represents the receiving node, and c represents the current node. Nodes include sending queues, receiving queues, sending power, growth rate, and other attributes. The growth rate is expressed as $\lambda \in \{0, 1\}$, it represents the change of the number of packets on nodes.

2.3 Optimization Objectives

The goal is to minimise the total time delay for transferring data depending on the network state. How to select the route, i.e., which node is the next packet hop to different nodes, can be summarised as a mathematical agent action selection problem. In this interpretation, the node plays the role of an agent, the packet route can be represented by the node action, and the channel quality can be expressed as the edge weight.

We transformed the maritime Ad Hoc network packet routing problem into a multi-agent behaviour selection problem. Corresponding to the multi-agent approach, we use $s \in S$ to represent the state set of adjacent nodes, s represents a specific state, $a \in A$ represents a limited action set, and a represents a specific action. Let $T(S, a, S') \sim P_r(S, a, S')$ be the agent transition model which predicts the next state s' based on the current state S and action a, where the P_r represents the probability of taking action a from s to s'; $R(s, a) = E[R_{t+1} \,|s, a]$ be an immediate reward for an action taken by an agent.

2.4 Problem Formulation

In this section, we propose a formula for the time delay minimisation problem based on reinforcement learning. A certain agent behavioural strategy leads to a positive reward in the environment, and then the tendency of the agent to enact this behavioural strategy in the future will be strengthened [4]. The agent's goal is to discover the optimal strategy in each discrete state to maximise the desired discount reward. We assume that the source domain is $U_A = \{(x_i^A, y_i^A)\}_{i=1}^{M_A}$, and the target domain is $U_B = \{(x_i^B, y_i^B)\}_{i=1}^{M_A}$, D_A and D_B are the hidden special invariants between the source domain and the target domain respectively. We define the classification function of the target domain as:

$$\psi(d_i^A) = \frac{1}{L_A} \sum_{j}^{L_B} y_i^B d_i^B (d_i^A)' = \Phi^B \Omega(d_i^A) \tag{1}$$

The objective function is shown as follows:

$$\arg \min_{\Theta^A, \Theta^B} L_1 = \sum_{i}^{M_c} l_1(y_i^B, \Psi(d_i^A)) \tag{2}$$

$$\arg \min_{\Theta^A, \Theta^B} L_2 = \sum_{i}^{M_{AB}} l_2(d_i^B, d_i^B) \tag{3}$$

The overall objective function is shown as follows:

$$\arg \min_{\Theta^A, \Theta^B} L = L_1 + \gamma L_2 + \frac{\lambda}{2}(\left\|\Theta^A\right\|^2 + \left\|\Theta^B\right\|^2) \tag{4}$$

3 Proposed Algorithms

To achieve efficient route allocation with lower time delays, isolated routing problems are transformed into multi-agent cooperative optimisation problems. We propose a federated reinforcement learning algorithm, which attaches a federated learning mechanism with similar nodes to reinforcement learning.

3.1 Motivation for Algorithm

In order to minimise the total packet forwarding process time, i.e., the waiting time plus transfer time, it is necessary to make optimal routing decisions based on the observations of surrounding nodes. Considering the policy similarity of neighbouring nodes, we used federated reinforcement learning to schedule the next hop packet selection.

The traditional centralized routing decision algorithm is not suitable for this scenario, especially when the number of packets is large. Another scenario is that centralized dispatching can lead to significant wait times when the packet is in an area where communication is poor. Based on the above problems, we consider to use a distributed routing decision algorithm. Meanwhile, since this problem has many influencing factors and is entangled with each other, it is not convenient to solve it in an analytical way, so we use the method of federated reinforcement learning to solve it. Intelligent routing algorithm based on reinforcement learning is able to handle higher dimensions of state characteristic information network, adaptive to different application scenarios and changes in the network environment, the reinforcement learning model and gives the intelligent routing algorithm not only focus on the current routing effect, more predictable future network status changes, and in advance to avoid network congestion what might happen in the future.

3.2 The Learning Common Policy Features of Similar Nodes

In an maritime Ad Hoc network, similar nodes have similar data and routing policies. They are expected to improve the inference accuracy of the model through joint learning. We cannot just apply federated learning to both sides of the data because the routing policies of different nodes are different. Both parties establish a reinforcement learning routing decision model, which have been recognised by their users in data acquisition. The problem is then how to establish high-quality models at each terminal. Due to incomplete or insufficient data, the reinforcement learning model at each end may not be established or lacks the ideal effect. Federated reinforcement learning can solve this problem by ensuring that the data of each node does not go out locally, allowing the federated system to optimise the learning model of all parties through an encrypted parameter exchange. However, when creating virtual models, the data itself does not move, nor does it compromise privacy or affect data compliance. Consequently, the constructed models serve only local goals within their respective regions.

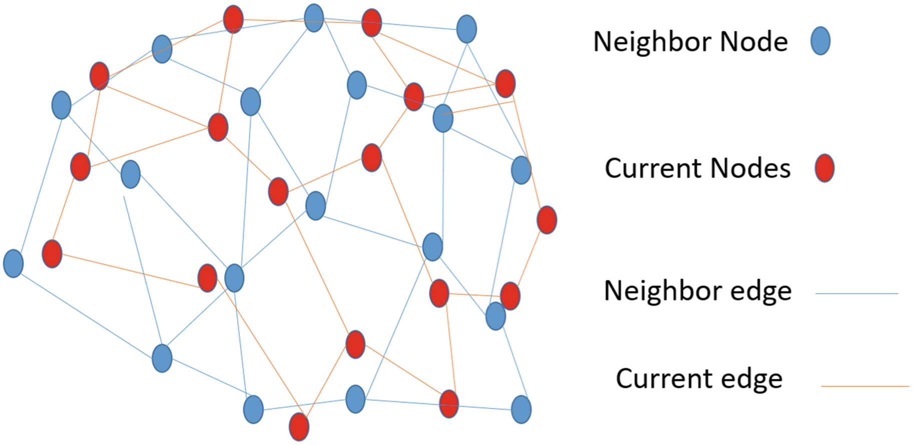

Fig. 2. Node association learning

Partition neighbor path planning based on federated learning focuses on how to map the data of neighbor nodes and current nodes from the original feature space to the new feature space. In this way, the data distribution of the base neighbor node is roughly the same as that of the current node, so that the labeled data samples of the base neighbor can be better used for classification training in the new space, and finally the data of the current node can be classified. To this end, we carry out feature mapping of nodes with close distance, so that neighbor nodes can be used to guide the model parameters of joint nodes with the trained model. Of course, there should be some structural similarity between the topology diagram of neighbor nodes and the current node. As shown in Fig. 2, we first train the neural network according to the red node data, and then take the trained neural network as the alternative network of the actual node. When new nodes join, or the data packet transmission rule of the current network changes, for example, the blue node and red topology are updated online by using federated learning method.

The reinforcement signal provided by the environment in federated reinforcement learning is an evaluation (usually a scalar signal) of the action generated by the agent, rather than telling the agent how to generate the correct action. Since the external environment provides little information, the agent must learn with similar nodes. Therefore, agents gain knowledge in an action-by-action evaluation environment and improve action plans to adapt to the environment. The aim of the reinforcement learning system is to dynamically adjust the parameters to achieve the maximum reinforcement signal. As the reinforcement signal R and the action a generated by the agent do not have a clear functional description, the gradient information R/a cannot be obtained. Therefore, in the reinforcement learning system, a random unit is needed. With this random unit, the agent will search in the possible action space and find the correct action.

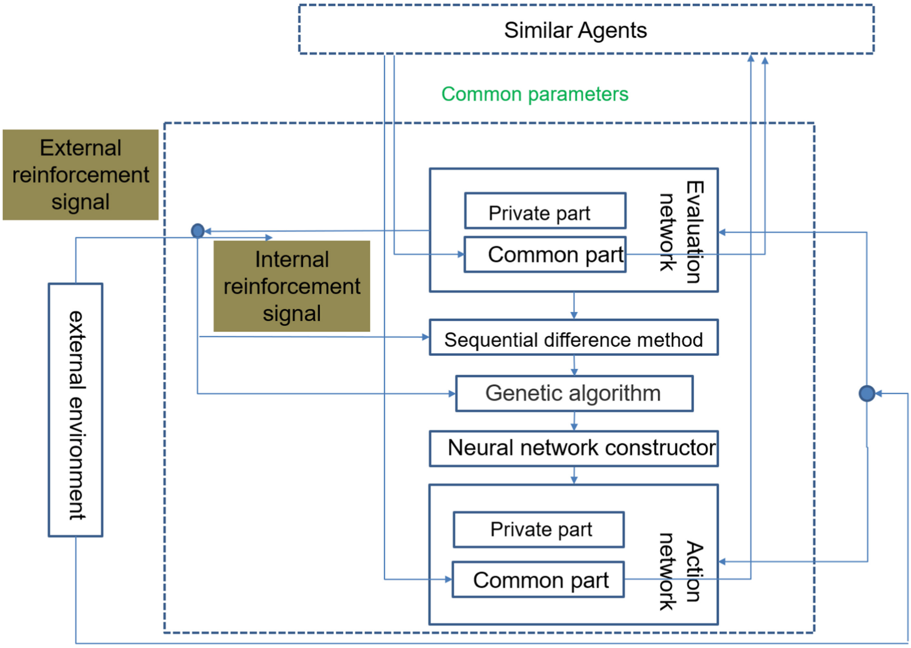

Fig. 3. An illustration of association learning

3.3 Cooperative Scheduling Mechanism Based on Transmission Task Completion

In reinforcement learning, the target of an agent is formally represented as a special signal, called reward, which is transmitted to the agent through the environment. At each time, reward is a single scalar value. Informally, an agent's goal is to maximize the total reward it receives. This means that it's not the immediate rewards that need to be maximized, but the cumulative rewards that need to be maximized over time. The use of reward signals to formalize goals is one of the most distinctive features of reinforcement learning.

The multi-agent path planning algorithm designed in this paper introduces the design reward of transmission task completion to carry out cooperative optimization under the framework of reinforcement learning, as demonstrated in Fig. 3. The principle of cooperative optimization algorithm is to decompose a complex objective function into simple sub-objective functions, and then carry out cooperative optimization of these sub-objective functions. Specifically, collaborative optimization is to optimize each sub-objective function while considering the results of other sub-objective functions, so that the optimization results among sub-objective functions can be consistent. The consistency of optimization results means that the values of each variable can be consistent in the optimization results of each sub-objective function.

The completion degree of this task represents the completion degree of transmission, and the feedback of task execution takes the difference between decision-making route and baseline route as reference. Effect prediction action coordination is mainly responsible for interaction eigenvalues of interested agents within the communication range. The information exchange of task completion is helpful for Agent coordination and strategy formulation in real scenes, and the interactive environment map information is helpful for a single Agent to execute decisions and avoid falling into local optimal solutions.

In this architecture, target behavior is learned from downstream task-specific rewards without any communication oversight. However, complex real-world tasks may need to take into account the interaction of agents after they complete their actions, such as the occurrence of congestion. Therefore, this capability needs to be enhanced by using a multi-round communication method, through which agents coordinate before taking action on the environment. First of all, each agent wants to transmit its own expected action and other agents accept the expected action of other agents at the same time. Then, according to the expected action of other agents, it changes its own action through the expected return and makes the real action. The agent then interacts with the real action environment. The state transition function of the decision is given by:

$$\begin{aligned} p(s_n{}', a'|s_{\,n}, a) = Pr(&s_{n+1,t} = s_n{}', A_{n+1,t} \\ &= a', R_{n+1,t} = a'|s_{\,n,t} = s_n, A_{n,t} = a) \end{aligned} \tag{5}$$

$$p(s_n{}', r|s_{\,n}, a) = Pr(s_{t+1,t} = s, R_{t+1,t} = a'|s_{\,t} = s, A_t = a) \tag{6}$$

3.4 Common Network Parameter Aggregation Methods

Each neural network is composed of two modules, namely a private network module and a common network module. In a private network, the federated reinforcement learning algorithm allows it to retain the private features. With the adjacent nodes' features from the common network, the action network output nodes can effectively complete a random search and greatly improve the possibility of selecting suitable actions. Furthermore, the entire action network can be trained online. With auxiliary network environment modelling, evaluation of networks based on the current status and external reinforcement signal simulation environment is used to predict a scalar value. This allows one step, and multi-step, prediction by the action network current actions to strengthen the signal applied to the environment, advance to the relevant action network to provide the candidate actions of intensive signals, and provide more information on rewards and punishments (internal reinforcement signal) [1]. This reduces uncertainty and speeds up learning.

The network operation is divided into two parts: reinforcement feedback calculation and joint parameter calculation. In reinforcement feedback calculation, the time-series differential prediction method (TD) and back-propagation algorithm (BP) are used to learn the evaluation network whilst genetic operation of the mobile network is conducted, and the internal reinforcement signal is used as the mobile network fitness function. Joint parameter calculation determines the weighted average of the parameters of similar nodes so that they can learn from each other. The private network provides more effective internal reinforcement signals to the mobile network, compelling it to produce more appropriate actions. The common network signals enable both the mobile and evaluation networks to learn together with similar nodes, thus greatly accelerating the learning of the two networks.

4 Performance Evaluation

Experimental Setup. The connections between nodes represent specific channels. When multiple data packets are transmitted on the network, they become congested at important nodes, which seriously affects the transmission capability of the entire system. We used federated reinforcement learning to make routing decisions and plan the routing choices of each packet at different nodes.

Simulation Results and Analysis. To simplify the simulation, we assume that the order of packets in the transmission queue does not change. Therefore, if the current packet is blocked, all subsequent packets will be blocked. To ensure that the total number of packets in the network will not exceed the upper limit, when the number of packets reaches the upper limit, one packet will be generated for every delivered packet. The packet generation rule $p^{i,j,k}$ is as follows:

$$p^{i,j,k} = p^{i,i,k} \; when \; p^{i,j,k} = p^{i,k,k} \tag{7}$$

$$i, k = random(0, n) \tag{8}$$

In the simulation we adopted this method to solve the maritime Ad Hoc network routing decision problem. To demonstrate the advantages of the FRL method, we chose to use the shortest path algorithm and Q-learning method for the simulation. The shortest path algorithm is a commonly used algorithm in the field of routing planning. The shortest path problem is a classical algorithm problem in graph theory, which aims to find the shortest path between two nodes in a graph. The learning algorithm allows the system to select the optimal action set by using the experienced action sequence in the Markov environment.

In Fig. 4, we depict the average delivery time versus the number of packets. Average delivery time is the time it takes for a packet to travel from its source to its destination. The number of packets was gradually increased from 500 to 5000, to study the effect of packet density. The trend of the points in the figure shows

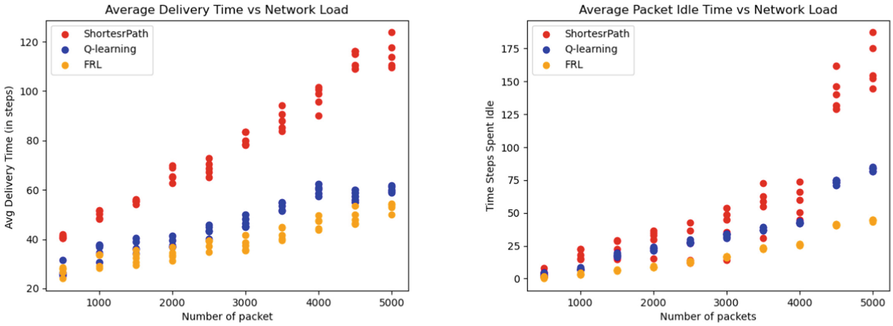

Fig. 4. Simulation results

that the average delivery time increases with packet density. The FRL algorithm has a slightly better performance than the Q-learning algorithm and is clearly better than the shortest path algorithm, thus reflecting the superiority of the algorithm. The relationship between the number of packets and the average packet idle time is shown in the Fig. 4. It can be seen that the FRL algorithm performs better in terms of average packet idle time. Therefore, nodes using the FRL algorithm have superior scheduling ability and avoid long idle packet times.

Through simulation, it was verified that the FRL algorithm can better solve packet congestion, ensure the speed of network transmission and make full use of node performance to avoid long packet idle times.

5 Conclusion

This paper investigated the distributed routing planning problem in maritime Ad Hoc networks with rapid topology changes and limited network bandwidth. With the aim of maximising throughput, the problem of transferring data efficiently was transformed into a congestion avoidance problem. Considering the differences in network nodes, the FRL is proposed to improve the efficiency of distributed routing planning through joint learning of similar nodes. Based on the dynamic data of the dedicated communication simulation system, the simulation results verify the performance of our method. In future work, we will study the application of FRL in private networks.

Acknowledgements. Key Projects of the National Defense Foundation Strengthening Plan under Grant 2020-JCJQ-ZD-020-05, National Defense Technology Innovation Research Institute under Grant 2020KCYWXGK4098, Talent Program of Liaoning Province under Grant XLYC1807149, National Key R&D Program of China under Grant 2020YFB1806800.

References

1. Chen, X., Yuan, Y., Lu, L., Yang, J.: A multidimensional trust evaluation framework for online social networks based on machine learning. IEEE Access **7**, 175499–175513 (2019)
2. Entezari-Maleki, R., Gharib, M., Rezaei, S., Trivedi, K.S., Movaghar, A.: Modeling and evaluation of multi-hop wireless networks using SRNS. IEEE Trans. Netw. Sci. Eng. **8**(1), 662–679 (2021)
3. Hanawal, M.K., Hayel, Y., Zhu, Q.: Effective utilization of licensed and unlicensed spectrum in large scale ad hoc networks. IEEE Trans. Cogn. Commun. Netw. **6**(2), 618–630 (2020)
4. Hwang, K.S., Jiang, W.C., Chen, Y.J., Hwang, I.: Model learning for multistep backward prediction in dyna-q learning. IEEE Trans. Syst. Man Cybern. Syst. **48**(9), 1470–1481 (2018)
5. Kim, B.S., Kim, K.I., Roh, B., Choi, H.: Hierarchical routing for unmanned aerial vehicle relayed tactical ad hoc networks. In: Proceedings of International Conference on Mobile Ad Hoc and Sensor Systems, pp. 153–154 (2018)
6. Liu, J., Guo, S., Shi, Y., Feng, L., Wang, C.: Decentralized caching framework toward edge network based on blockchain. IEEE Internet Things J. **7**(9), 9158–9174 (2020)
7. Mowla, N.I., Tran, N.H., Doh, I., Chae, K.: Federated learning-based cognitive detection of jamming attack in flying ad-hoc network. IEEE Access **8**, 4338–4350 (2020)
8. Naseer Qureshi, K., Bashir, F., Iqbal, S.: Cloud computing model for vehicular ad hoc networks. In: Proceedings of IEEE International Conference on Cloud Networking, pp. 1–3 (2018)
9. Peng, J., Li, X., Li, X.: Research on election interval of distributed wireless ad hoc networks. IEEE Access **8**, 110164–110171 (2020)
10. Ramli, N.I.S., Hisham, S.I., Ismail, N.S.N., Ramalingam, M.: Performance comparison between AODV and DSR in mobile ad-hoc network (MANET). In: Proceedings of IEEE ICSECS-ICOCSIM, pp. 217–221 (2021)
11. Rukaiya, Khan, S.A.: Self-forming multiple sub-nets based protocol for tactical networks consisting of SDRS. IEEE Access **8**, 88042–88059 (2020)
12. Sattler, F., Wiedemann, S., Mller, K.R., Samek, W.: Robust and communication-efficient federated learning from non-I.I.D. data. IEEE Trans. Neural Netw. Learn. Syst. **31**(9), 3400–3413 (2020)
13. Shi, Y., Li, W., Zeng, W.: A study on interaction of college English classroom in the mobile internet environment. In: Proceedings of IEEE IWCMC, pp. 1766–1769 (2021)