A Diversified Message Type Forwarding Strategy Based on Reinforcement Learning in VANET

Guoai Xu[1,3], Boya Liu[1,2,3], Guosheng Xu[1,3*], Peiliang Zuo[2]
[1]School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China
[2]Beijing Electronic Science and Technology Institute, Beijing, China
[3]National Engineering Laboratory of Mobile Network Security, Beijing University of Posts and Telecommunications, Beijing, China
[E-mail: guoshengxu@bupt.edu.cn]
[*]Corresponding author: Guosheng Xu

Received December 18, 2021; revised March 20, 2022; revised July 30, 2022; accepted September 1, 2022; published September 30, 2022

Abstract

The development of Vehicular Ad hoc Network (VANET) has greatly improved the efficiency and safety of social transportation, and the routing strategy for VANET has also received high attention from both academia and industry. However, studies on dynamic matching of routing policies with the message types of VANET are in short supply, which affects the operational efficiency and security of VANET to a certain extent. This paper studies the message types in VANET and fully considers the urgency and reliability requirements of message forwarding under various types. Based on the diversified types of messages to be transmitted, and taking the diversified message forwarding strategies suitable for VANET scenarios as behavioral candidates, an adaptive routing method for the VANET message types based on reinforcement learning (RL) is proposed. The key parameters of the method, such as state, action and reward, are reasonably designed. Simulation and analysis show that the proposed method could converge quickly, and the comprehensive performance of the proposed method is obviously better than the comparison methods in terms of timeliness and reliability.

Keywords : Routing strategy, Message Forwarding Strategy, Reinforcement Learning, VANET, Black nodes

This research was funded by the National Key Research and Development Program of China (No.:2021YFB3101500)

http://doi.org/10.3837/tiis.2022.09.015 **ISSN : 1976-7277**

1. Introduction

As of the beginning of 2020, 330 million vehicles over the world have been interconnected [1]. VANET provides users with safe, efficient, convenient and comfortable intelligent messaging services, which has become an important part of modern intelligent transportation and has been closely watched by experts and scholars at home and abroad [2]. And among the many research fields of VANET, routing is one of the key technologies that restrict its development. So many researchers have attempted to optimize routing methods suitable for VANET over the years.

Greed Perimeter Stateless Routing (GPSR) algorithm is a typical location-based routing protocol [3]. The nodes in this protocol do not need to maintain routing tables and are simple and easy to implement. It has become a common routing protocol for VANET systems. The application of reinforcement learning [4] algorithm to the Internet of vehicles can optimize the quality of service parameters in VANET to different degrees, such as delay, bandwidth, security, etc. This provides a better solution for the design and improvement of routing strategy in VANET. Reinforcement learning is an important field of artificial intelligence, in which the main research problem is that the agent should obtain the maximum reward value through interaction with the environment, and thus learn the strategy with the best return. Since the state-action pair matching problem in reinforcement learning is similar to the dynamic routing process, it has been widely used in this respect, which also provides a good solution for the design and improvement of message forwarding strategies in VANET.

J. Li [5] proposed a Q-learning-based routing algorithm for wireless sensor networks. Each node is marked as a state, and the transition of the state is defined as an action. The optimal path is established by traversing the routing table. M. Yuan [6] et al. raised a multi-priority message-oriented VANET routing algorithm based on the Q-value, which aims to solve the problem of load balancing. By considering location information and received signal strength, the routing problem is transformed into a Q-learning optimization process on the basis of fuzzy constraints in [7]. C. Wu et al. [8] proposed an improved Q-learning routing protocol QLAODV, which could effectively deal with the high-speed dynamic movement of mobile ad hoc networks and frequent topology changes. R. Plate et al. [9] proposed a Q-learning routing method QKS that combines kinematics and scanning features to solve the problem of slow convergence caused by the Q-learning algorithm. In order to solve the multicast problem and improve the performance of the MAODV routing protocol, G. Santhi et al. [10] proposed the MANET multicast routing protocol QLMAODV by applying the Q-learning algorithm to the existing MAODV protocol. By preemptively selecting the sub-optimal routing before the failure of the current active routing for network state learning, Y. Sun et al. [11] proposed a location-based reinforcement learning routing protocol PBQR, which defines the stability factor and continuity factor, uses the Q-learning algorithm to evaluate the quality of neighbor nodes, selects the next hop node based on the node location information, and enhances link stability and reliability. J. Wu et al. [12] proposed an adaptive routing protocol based on reinforcement learning (ARPRL). By designing a new Q-value update function, using the DATA forwarding mechanism and the MAC layer feedback mechanism to assist in updating the Q-value table, it effectively solves the routing loop, link interruption and other issues. S. Jiang et al. [13] designed an auxiliary geographic routing based on Q-learning to improve the performance of data packet transmission and end-to-end delay. J. Aznar-Poveda et al. [14] proposed a joint beacon rate and transmission power control based on Q-learning and policy evaluation to ensure the timeliness of message forwarding.

Currently, most of the current research work focuses on algorithm innovation and improvement under the premise of a certain fixed goal. However, faced with the fact that the increasingly complex needs of users which leads to the diversity of network forwarding messages, designing a message forwarding strategy that selects different forwarding protocols for different message types, thereby reducing network overhead, is a key direction in the field of vehicular networking research. To the best of the authors' knowledge, there is quite little research on the routing methods related to the message types of VANET. In the "Web 5.0 Technology" white paper released in September 2021 [15], internet information is divided into 10 levels of certainty. This paper defines the messages into four types take into account the timeliness and reliability requirements of message transmission process, and two routing algorithms with completely different transmission processes are introduced. RL algorithm has fast convergence speed when the model is not too complex and can meet the requirement of low delay of scheme model. For this reason, we use reinforcement learning algorithm to achieve intelligent transmission strategy matching with different message types.

The follow-up content of the paper is arranged as follows. Section 1 introduces the related preliminary knowledge. Section 2 describes the system model in detail. Section 3 covers algorithm modeling combined with reinforcement learning and gives the details of the proposed algorithm. Section 4 verifies the specific performance of the method proposed in this paper. The last section summarizes the paper.

2. Preliminary knowledge

2.1 GPSR

Location-based protocols is a promising routing solution for VANET, regarding their performance. GPSR is a typical location-based routing algorithm. The node obtains the location information of the node through the positioning system, and uses the greedy forwarding algorithm to select the closest node to the destination node within the communication range to forward the message. If there is no communicable node, the neighboring forwarding algorithm is used to avoid routing holes. The node uses greedy forwarding and peripheral forwarding according to the situation until the communication is completed.

Until now, many scholars have made improvements to the GPSR protocol [16-19], especially to make it applicable to VANET. H. Yuan et al. [20] proposed GPSR-TM protocol. In this protocol, the relationship between vehicles is expressed in the form of social network, and the location, speed and social attributes of neighbor nodes are considered when the next-hop forwarding node is considered. Wang et al. [21] proposed a GPSR algorithm based on prediction. In each broadcast, the location of neighbor nodes is predicted by considering the node speed and direction, and then the next hop node is selected. A. Benmir et al. [22] proposed to send the same data packet in two different paths to maximize its receiving probability. Simulation results show that the scheme is better than GPSR in some performance. S. Younes et al. [23] proposed a novel Extended Kalman Filter Greedy Perimeter Stateless Routing protocol. EKF-GPSR uses a stochastic prediction model based on an EKF to obtain the user information when transmitting data instead of using information from the last beacon exchanged messages which is most likely outdated.

2.2 Reinforcement learning

Reinforcement learning is a kind of learning that maps the action space from the state space to enable the agent to obtain the greatest reward in the process of interacting with the environment [24]. One of the most important algorithms of RL is Q-learning that involves action-value function $Q(a,s)$ which refers to the expected reward when taking a pair of action a a and state s. It is defined as:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a') - Q(s,a)) \quad (1)$$

The Markov decision process can usually be used to model reinforcement learning. $\varepsilon - greedy$ is a way to reach a better compromise between exploration and utilization, and it can be described as:

$$a = \begin{cases} randomly \quad action \quad with \quad probability \quad \varepsilon \\ \arg\max_{a \in A} Q(a,s) \quad with \quad probability \quad 1-\varepsilon \end{cases} \quad (2)$$

The application model in this paper is single-step reinforcement learning, which corresponds to the theoretical model of the "K-rocker gambling machine" [25].

2.3 Dempster-Shafer (D-S) evidence theory

Dempster-Shafer (D-S) evidence theory, proposed by Harvard University mathematician A.P. Dempster, was further improved by his student Shafer [26]. D-S evidence theory is a complete theory to deal with the uncertainty problem. It can not only emphasize the objectivity of things, but also emphasize the subjectivity of human estimation of things. Its biggest feature is that the description of uncertain information uses "interval estimation" instead of "point estimation". It distinguishes between unknown and uncertainty, and shows great flexibility in accurate reflection of evidence collection. It is often defined as follows:

(1) Recognition frame: the commonly used symbol Θ stands for recognition frame, which means all possible answers to a certain question, but only one answer is correct.

(2) Mass function: use $m()$ to represent the mass function, which reflects the degree of trust. Among them, $m: 2^{\Theta} \to [0,1]$.

(3) Trust function: $Bel(A)$ is defined as the sum of the basic probability distributions of all subsets of A, which represents all trust in A, that is: $Bel: 2^{\Theta} \to [0,1]$, $Bel(A) = \sum_{B \le A} m(B) = 1, A \subseteq \Theta$.

(4) Likelihood function: $Pl(A)$ means that the trust degree of A is not denied, and it is the sum of the basic probability distributions of all the subsets that intersect with A. The uncertainty measure that A seems to hold can be expressed as: $Pl: 2^{\Theta} \to [0,1]$, $Pl(A) = 1 - Bel(\overline{A}), A \subseteq \Theta$.

3. System model

The routing scenario covered in this paper is shown in **Fig. 1**. It's worth noting that, due to the vulnerability of wireless communication, the adversary will launch various attacks against routing. Malicious nodes have various malicious behaviors. They may execute eavesdropping attack, denial of service attack, impersonation attack, black hole attack and so on [27]. Malicious nodes basically disguise themselves as normal nodes to participate in node forwarding. After messages are transmitted to malicious nodes, malicious will analysis and use the messages to destroy the network or gain illegal benefits. The behavior of

malicious nodes mainly increases the end-to-end transmission delay and reduces the overall network performance in terms of its impact on the performance of VANET. In order to simulate the real scene as much as possible, we added malicious nodes into the system model. In this paper, malicious nodes specifically refer to "black hole" nodes [28]. This assumption will make the description of the system model more intuitive. The behavior of the black hole node can be expressed as after receiving a message, the node will discard the message instead of forwarding it to the next relay node. Malicious nodes will increase end-to-end transmission delay, cause waste of network resources, and reduce the performance of the entire network.

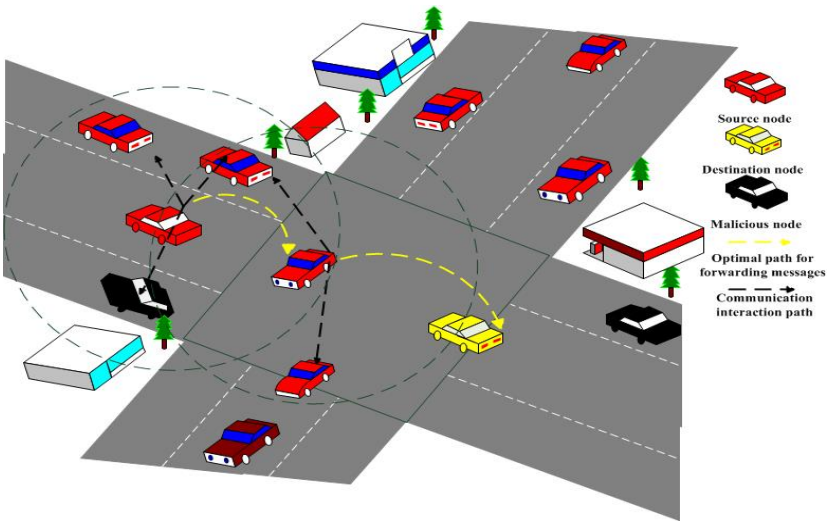

Fig. 1. Schematic diagram of message forwarding strategy scenario

We assume the communication range of vehicle nodes is limited. So nodes need to cooperate with each other to complete data forwarding. Information such as position, speed and trust value of vehicle nodes within the communication range can be obtained periodically, and the message can be transmitted to the destination node through relay nodes. In the figure, the source node (vehicle) selects a matching forwarding action according to the forwarding message type, and transmits the message to the relay node. The relay node should be selected according to the message type, node distance, speed, trust value and other factors. The relay nodes participating in the forwarding cycle execute the policy to complete the communication process. After the communication is completed, the trust value and Q-value of different actions of corresponding nodes participating in the communication process can be updated simultaneously. For ease of understanding, the calculation process of updating Q-value and trust value is shown in detail below.

For state s_i, the update process of the average reward value $Q_m(s_i, a_j)$ of the m-th attempt of action a_j is as follows:

$$Q_m(s_i, a_j) = \frac{1}{m}\left((m-1) \times Q_{m-1}(s_i, a_j) + \gamma_m\right) \tag{3}$$

Where state $s_i \in S$, S is the state space, action $a_j \in A$, A is the action space, γ_m represents the reward value obtained in the m-th attempt.

According to D-S theory, the comprehensive trust value of nodes includes objective trust value and subjective trust value. The objective trust value is the trust value obtained by direct communication with the source node, while the subjective trust value is the trust value obtained by communication between the node and its neighbor node. Each node will hold the trust value of other nodes. Since the integrated trust value needs to be calculated through the communication process when performing D-S element actions, the network overhead is larger than the GPSR algorithm. An example of the calculation process of the trust value is as follows. Assuming that the forwarding node n_1 has neighbor nodes n_2 and n_3. According to the D-S evidence theory, the objective trust value of n_2 and n_3 are the trust function $Bel(n_2) = mass(n_2)$ and $Bel(n_3) = mass(n_3)$. The subjective trust value is $Bel(\Theta) = mass(\Theta)$. Then the comprehensive trust value of the neighbor node can be obtained by the likelihood function.

$$Pl(n_2) = Bel(n_2) + Bel(\Theta) = mass(n_2) + mass(\Theta) \quad (4)$$

$$Pl(n_3) = Bel(n_3) + Bel(\Theta) = mass(n_3) + mass(\Theta) \quad (5)$$

Thus, n_1 gets the trust values $Pl(n_2)$ and $Pl(n_3)$ of n_2 and n_3.

4. QLMTR strategy model

This section will describe the Q-learning based message type routing (QLMTR) strategy/algorithm. To construct a strategy model based on Q-learning, state space, action space and reward function need to be established. The state space, action space, reward functions of the strategy proposed in this paper will be described in detail below.

4.1 State space

As a preliminary study of message forwarding strategy, in order to facilitate the expression of the method, we further simplify the message classification defined in [15]. Specially, message types are described and analyzed in accordance with the requirements of urgency and reliability, and state space S is defined as a set of message types accordingly. It is worth mentioning that the type and quantity of message types can be adjusted according to specific needs in practical applications. In this paper, S is classified as follows:

Type I messages is defined as the messages with emergency and high reliability requirements, which refers to the information that can directly affect the driving safety of the vehicle, such as the driving state of the vehicle itself. This information requires high timeliness and reliability, and cannot be lost during the transmission process. It requires the success of the transmission in t_1 times of communication interactions.

Type II messages correspond to the messages with emergency but low reliability requirements, which refers to information that can slightly affect the driving process of the vehicle, such as road surface information, road section information, etc. This information is required to be delivered to the vehicle node in real time. During the execution of the algorithm, the source node is required to try to send at most t_2 times.

Type III messages correspond to the messages with emergency but high reliability requirements, which refers to information that might affect the driving safety of vehicles, such as congestion, traffic density, and other information. This kind of information is required to be accurately transmitted to vehicle nodes. During the execution of the algorithm, the source node is required to try to send at most t_3 times.

Type IV messages is defined as the messages with non-urgent but low reliability requirements, which refers to news, information, entertainment and other information, that serves drivers but does not affect driving safety. During the execution of the algorithm, the source node is required to try to send at most t_4 times.

Considering the timeliness requirements of different message types, in general, $t_1 \le t_2 \le t_3 \le t_4$.

4.2 Action space

The meta-action includes the message transmission based on the GPSR routing algorithm (hereinafter referred to as GPSR meta-action), and the message transmission of the improved GPSR routing algorithm based on the D-S evidence theory (hereinafter referred to as the D-S meta-action), both alone and a combined action of the two compose the action space A.

We designed it in this way because GPSR is a typical routing protocol. Similarly, D-S evidence theory has been proved to have a good effect on trust measurement. In order to illustrate the methods and performance of the intelligent routing strategy proposed in this paper, we select these two typical methods as meta-actions. It can be pointed out that the action space can choose different routing methods according to actual needs.

4.2.1 Meta-action

GPSR meta-action. The message transmission action is executed according to the GPSR standard algorithm.

D-S meta-action. The D-S meta-action uses trust measures to solve the situation that there may be malicious nodes in VANET. The source node uses the D-S evidence theory to measure the comprehensive trust value of the node, and selects the node with high comprehensive trust value as the forwarding node, and the forwarding node repeats this method until the message is transmitted to the destination node.

4.2.2 Combined action

The number of actions in the action space A is related to the number of transmission attempts by the source node. This paper takes t_2 transmissions required for Type II messages as an example. For the convenience of explanation, let $t_2 = 2$, and the corresponding actions in the action space are shown in **Table 1**.

Table 1. Action combination table

	First transmission	**second transmission**
Action 1	GPSR meta-action	GPSR meta-action
Action 2	GPSR meta-action	D-S meta-action
Action 3	D-S meta-action	GPSR meta-action
Action 4	D-S meta-action	D-S meta-action

4.3 Reward function

As different message types have different requirements for forwarding actions, this paper set that the number of transmissions and the transmission success rate respectively correspond to the requirements of transmission urgency and reliability, for instance, it is not necessary to consider their network overhead for the message type with high timeliness and reliability requirements. The reward and punishment functions of the four types of messages are as follows.

Since urgent and reliable (Type I) messages require a small number of transmissions and a high transmission success rate, the reward value function of this state is defined as $\gamma = \left(\frac{\alpha_1}{t} + \beta_1 \right) R$.

The two message states, which are urgent but have low reliability requirements (Type II), and not urgent but have low reliability requirements (Type IV), do not require high transmission success rate. The reward value function of this state is defined as :

$$\gamma = \left(\frac{\alpha_2}{t_{GPSR} C_{GPSR} + t_{D-S} C_{D-S}} + \beta_2 \right) R \tag{6}$$

For a message state that is not urgent but requires high reliability (Type III), the number of transmissions can be relatively high, so the reward value function in this state is defined as:

$$\gamma = \left(\frac{\alpha_3}{t_{GPSR} C_{GPSR} + t_{D-S} C_{D-S}} + \beta_3 \right) R \tag{7}$$

Among them, R is the indicator function, when the transmission is successful, $R = 1$, and $R = 0$ if unsuccessful. Both $\alpha_{(\cdot)}$ and $\beta_{(\cdot)}$ are weighting factors, which respectively represent the number of transmissions and the proportion of transmission success in the reward, $\alpha_{(\cdot)} > 0$, $\beta_{(\cdot)} \geq 0$, and $\alpha_{(\cdot)} + \beta_{(\cdot)} = 1$. t is the number of transmissions, t_{GPSR} and t_{D-S} are the number of transmissions using GPSR meta-action and D-S meta-action in t transmission, $t_{GPSR} + t_{D-S} = t$. C_{GPSR} and C_{D-S} are the network overheads using GPSR meta-action and D-S meta-action respectively, generally, $C_{D-S} \geq C_{GPSR}$.

4.4 Algorithm Details

The specific process of the forwarding strategy mentioned in this paper is shown in **Algorithm 1.** In this paper, we assume that each vehicle is equipped with GPS position module. Vehicles can obtain the location, speed, trust value and other information of the neighbor node. The reward value is calculated according to the reward function for the selected action corresponding to the type of message forwarded each time. The calculation method of trust value is based on D-S theory, please refer to the section of system model. We set R as the node communication radius, and L is the distance between nodes.

Algorithm 1. Details of the QLMTR strategy

Initialize source node p_1 and destination node p_2 .

p_1 periodically sends and receives Hello packets to obtain neighbor nodes within the communication radius R .

If $L_{p_1 p_2} \leq R$ **, then** the source node directly completes the communication with the destination node.

Else $L_{p_1 p_2} > R$ **, then** p_1 according to the type of message forwarded, select the corresponding action and update the reward Q-value.

If p_1 selects GPRS meta-action, **then** the GPSR routing algorithm is directly executed to complete the communication.

Else p_1 selects D-S meta-action, **then calculates** the trust value of each neighbor node, and executes the improved GPSR algorithm based on D-S to complete the communication.

End If

End If

p_1 and p_2 update their trust values for the nodes participating in the forwarding process. At the same time, p_1 updates the reward value Q function with reference to the reward function corresponding to the current state.

5. Simulation and analysis

To simplify the description, we only show the simulation results of Type I messages, Type III messages and Type IV messages. The simulation parameter settings are shown in **Table 2**. In the simulation, the number of simulations for each message under the condition of the number of malicious nodes is 200 times. In order to highlight the huge difference between GPSR meta-action and D-S meta-action in terms of network overhead, they are set as 1 and 30 respectively. In order to highlight the reasonable effectiveness and advantages of the strategy proposed in the paper, the specific simulation results are shown below.

Table 2. Simulation parameter settings

parameter	value
Total number of nodes	100
Number of malicious nodes	2~50
α_1	0.5
β_1	0.5
α_2	0.9
β_2	0.1
α_3	0.1
β_3	0.9
t_1, t_2, t_3, t_4	1
C_{D-S}	30
t_1, t_2, t_3, t_4	1,2,3,4
Simulation area	1000m*1000m
Node transmission radius	250m
Trust threshold	0.7
Transmission Preparation delay $T_{prepare}$	10ms
GPSR meta-action T_{action}	2ms
D-S meta-action T_{action}	6ms

5.1 Simulation results for Type I messages

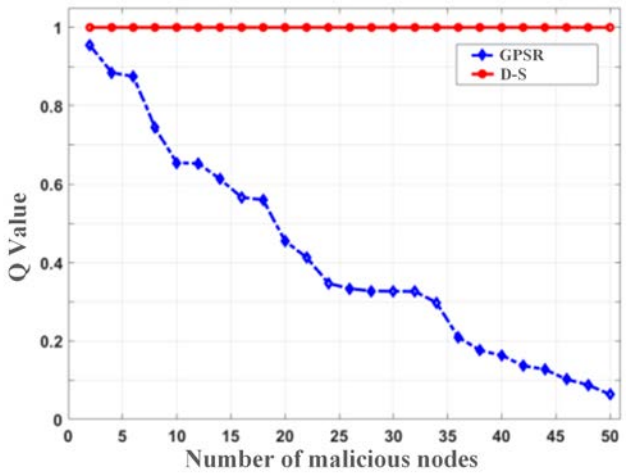

Fig. 2. Q-value of type I message meta-action VS Number of malicious nodes

The Q-value changes of the two meta-action of Type I messages under different numbers of malicious nodes are shown in **Fig. 2**. It can be clearly seen from the figure that for urgent and reliable messages, in the presence of malicious nodes, the sending node can accurately find the transmission method (i.e. D-S meta-action) that should be used through the strategy designed in this paper. In particular, it can also be seen that with the increase in the number of malicious nodes, the Q-value of GPSR meta-action shows a clear downward trend. In contrast, the Q-value of D-S meta-action has always remained the same. This is due to the fact that the transmission method based on the D-S evidence theory can accurately identify malicious nodes in VANET, and prevent such nodes from participating in the forwarding process of messages with high timeliness and reliability. At the same time, the transmission method based on GPSR algorithm is not with this capability, with the increase of malicious nodes, more failed transmissions will inevitably occur.

5.2 Simulation results for Type III messages

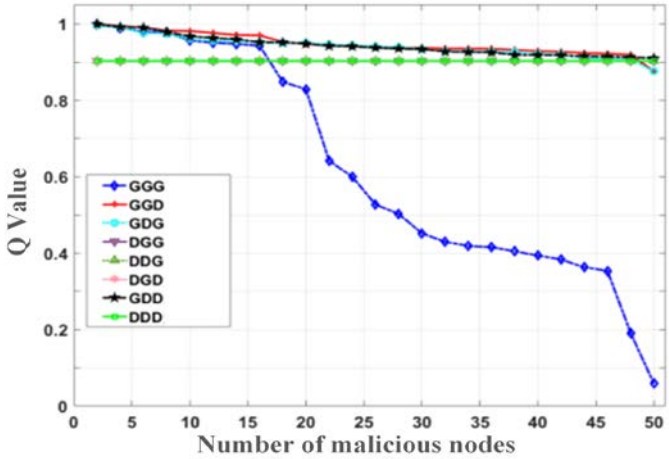

Fig. 3. Q-value of type III message meta-action VS Number of malicious nodes

For non-urgent and reliable messages, since the maximum number of transmissions allowed is 3, there are 8 meta-action to choose from. The Q-value changes of different meta-action combinations with the number of malicious nodes are shown in **Fig. 3**. It can be clearly seen from the figure that the Q-value of the GGG meta-action is higher when the number of malicious nodes is low (i.e. less than 17). As the number of malicious nodes increases, the Q-value decreases significantly. Meanwhile, it can be seen that the Q-value changes of DGG meta-action, DDG meta-action, DGD meta-action, and DDD meta-action are exactly the same. The same as the aforementioned reasons, they all use the method based on DS evidence theory when transmitting the message for the first time. So this method has a high success rate. Since the second and third transmissions are not required, the Q-value is exactly the same. In order to further compare the remaining three meta-actions (GGD meta-action, GDG meta-action, and GDD meta-action), **Fig. 4** shows the Q-value changes of the three ways. It can be seen from the figure that the Q-value of the GGD meta-action is higher than the other two meta-action in more than half of the scenarios, which shows that when the sending node uses the strategy proposed in this paper, it tends to use the GPSR method for transmission first, and use the D-S method for transmission in the last transmission. This is owing to that Type III messages are relatively insensitive to timeliness and allow transmission to fail, but require that the information must be successfully transmitted within the maximum number of transmissions. At the same time, it can be concluded that through the strategy proposed in this paper, the nodes in VANET can adaptively adjust the transmission method that should be adopted according to the network situation and their own message sending requirements. In order to achieve the transmission purpose of the condition, the overall benefit is the highest.

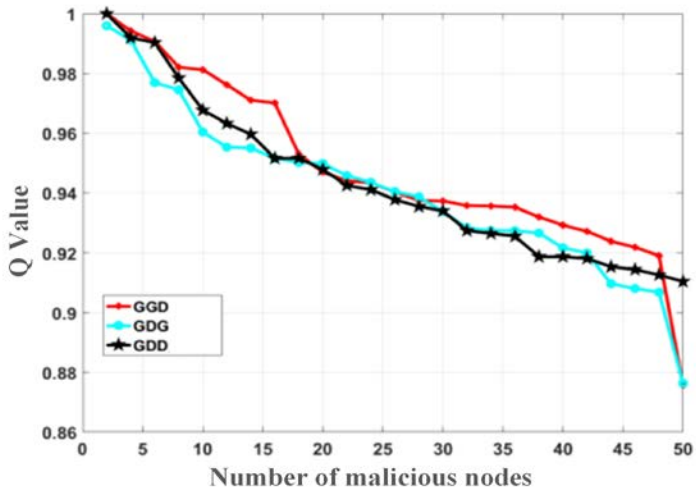

Fig. 4. Changes in the Q-value of GGD, GDG, and GDD in Type III messages as malicious nodes increase

5.3 Simulation results for Type IV messages

For Type IV messages, the maximum number of transmissions is set to 4, so the action space is composed of 16 actions. The Q-values of these actions under different numbers of malicious nodes are shown in **Fig. 4**. Similarly, the GGGG meta-action is most sensitive to the number of malicious nodes. When the malicious node exceeds 40, its Q-value is much lower than the other 15 meta-action. In addition, it can be seen that for the first transmission

of meta-actions (that is, DXXX-type meta-action) which use the transmission method based on the D-S evidence theory, their Q-values always remain unchanged and consistent.

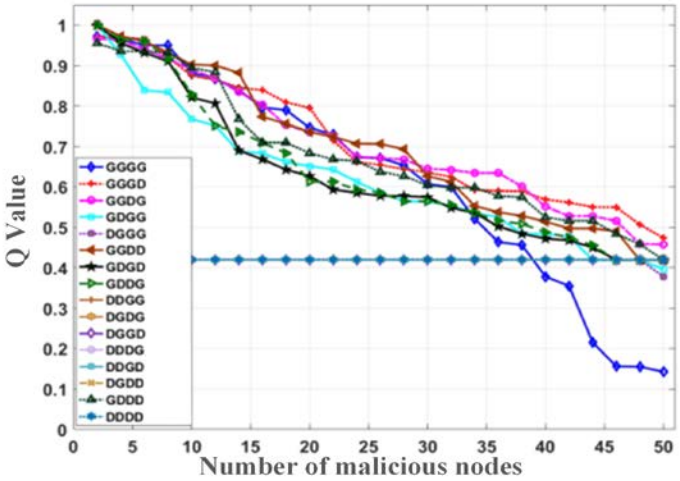

Fig. 5. Q-value of category IV message meta-action VS Number of malicious nodes

In order to further compare the remaining meta-action, **Table 3** shows the meta-action with the largest Q-value for different malicious nodes. Although none of the meta-actions shows absolute advantages, it is not difficult to observe from the table that it is allowed to perform In the case of four transmissions, the meta-action with the GPSR mode in the first two transmissions account for 96% of all the meta-action with the largest Q-value. This is basically consistent with the principle of the strategy proposed in this paper, that is, it is unreliable for non-urgency. For such messages, more attempts should be made to use the GPSR method with lower network overhead for transmission. However, when there are a large number of malicious nodes in the network, the GPSR method will have a high probability of transmission failure. In this case, the DS method should be used. What needs to be added is that the Q-values of the remaining types of meta- actions are not much different (see **Fig. 5**), and their overall Q-value changing trends remain the same, which shows that the meta-action selection of the sending node when sending Type IV messages is relatively flexible.

Table 3. Actions with the maximum Q-value for different malicious nodes in Type IV messages

Number of malicious nodes	2	4	6	8	10	12	14	16	18
Meta-action with the largest Q value	GDGG/GGDD	GGGD	GD GD	GD DG	GG DG	GG GD	GG DD	GD DD	GG DG
Number of malicious nodes	20	22	24	26	28	30	32	34	36
Meta-action with the largest Q value	GGGG	GGGD	GG GG	GG GD	GG GG	GG DD	GG DD	GG DG	GG GD
Number of malicious nodes	38	40	42	44	46	48	50		
Meta-action with the largest Q value	GGGD	GDDD	GG GD	GD DD	GG DG	GG GD	GG DG	/	

5.4 Comparison of transmission times

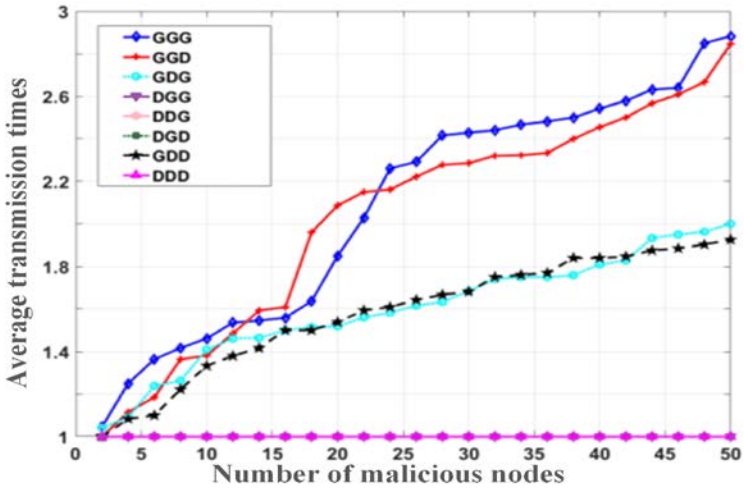

Fig. 6. The average number of transmissions of meta-actions in Type III messages

This paper uses Type III messages as an example to illustrate the comparison of transmission times. Note that the transmission times of other types of messages also have a similar trend. Due to the limitation of the length of the paper, this paper will omit the relevant content. **Fig. 6** shows the average number of transmissions for the eight types of meta-action combinations in the transmission of Type III messages. An interesting result can be observed, that is, when the sending node uses the DS method for the first transmission (that is, DXX-type meta-action), the total number of transmissions is always 1. The reason is consistent with the previous analysis, i.e., when the DS mode is used for transmission, the message can be successfully transmitted at one time. For the same reason, the second transmission is a D-S meta-action (GDG and GDD), and the maximum number of transmissions will not exceed two. As for the remaining GGG meta-action and GGD meta-action, it can be seen from the figure that the average transmission times of the two increase with the increase of malicious nodes. This is because the increase of malicious nodes reduces the transmission success rate of the GPSR method. Therefore, it is necessary to increase the number of transmissions in exchange for the transmission success rate.

5.5 Comparison of the number of transmission nodes

Similarly, only the average number of nodes participating in forwarding of different meta-action combinations of type III messages is used as an example to illustrate the correlation. The changes in the other three types of messages are similar. It can be seen from **Fig. 7** that as the number of malicious nodes increases, the average number of participating nodes for DGG meta-action, DDG meta-action, DGD meta-action, and DDD meta-action that perform only one transmission also increases. This is due to reliability. The reduction in the number of forwarding nodes affects the transmission path. Similarly, for the remaining four meta-action, their average number of participating nodes also shows a significant increase, and due to the influence of the number of transmissions, the number of participating nodes in GGG meta-action and GGD meta-action are also significantly greater than that of GDG meta-action and GGD meta-action GDD meta-action.

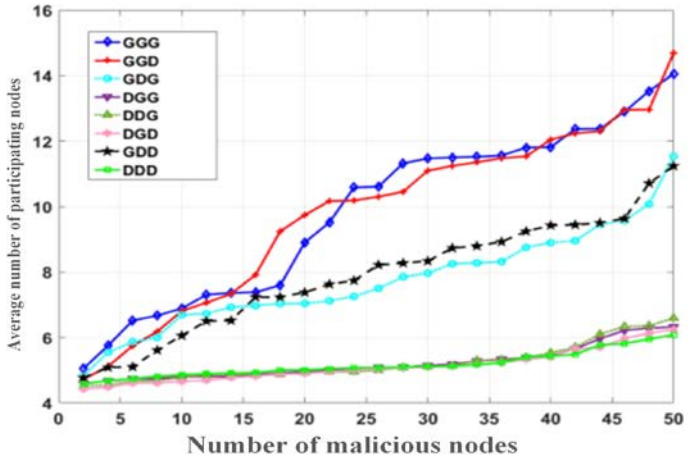

Fig. 7. The average number of participating nodes for meta-actions in Type III messages

5.6 Convergence

The strategy proposed in this paper is based on the theory of single-step reinforcement learning. It means that the action selection faced by the sending node and the forwarding node is a two-choice or multiple-choice process, so the process of strategy learning has the characteristics of low complexity and fast convergence. According to the simulation observation of convergence, it is found that for the four types of messages studied in this paper, the update process of the corresponding meta-action Q-value can reach convergence in the process of 200 message transmissions. Due to space limitations, this paper only specifically gives the convergence of type III messages under the condition of 16 malicious nodes. As shown in **Fig. 8**, the update process of the Q-value is obtained by averaging 1000 times. For other types of messages, the convergence process is similar. It can be clearly seen from the figure that for different meta-action combinations, owing to the reward and punishment functions in the respective learning process are different, there is a certain difference in the speed of convergence. However, different combinations of meta-action can converge to a stable Q-value in 58 message transmissions. Therefore, the strategy proposed in this paper has a faster convergence rate.

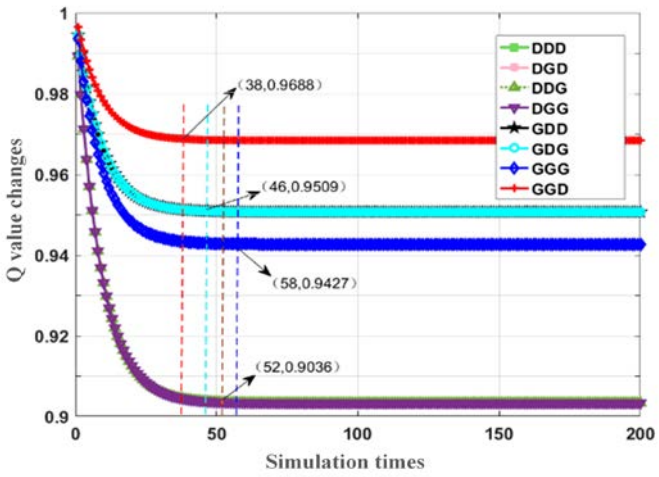

Fig. 8. Convergence of the Q-value learning process for the meta-actions of type III messages, with 16 malicious nodes

5.7 Strategic advantage

In view of the fact that there is no forwarding strategy for the diversity of message categories in the VANET scenario, this paper compares the GPSR algorithm and the improved GPSR algorithm based on D-S evidence theory from the reliability of message forwarding and network overhead, and compares the three transmission delay situation and complexity are compared specifically to reflect the advantages of the proposed strategy.

In the GPSR algorithm, a node obtains the status of n nodes around it and uses greedy forwarding to select the nearest node to the destination node to transmit messages. GPSR algorithm does not need to maintain the routing table, but it needs to planarization the network topology with a complexity of $O(n)$, where n is the density of neighbor nodes. For the improved GPSR algorithm based on D-S evidence theory, the node obtains the status of n nodes around it. The node will use D-S evidence theory to calculate node trust value, and select node with high trust value for message transmission. Due to the method only involves the superposition of trust values of nodes based on GPSR, its complexity is $O(n)$. In our paper, the node obtains the status of n nodes around it, and selects GPSR or improved GPSR transmission according to the type of messages to be transmitted. Furthermore, the general definition of the complexity of machine learning algorithms is mainly concerned with the complexity of their utilization, mainly reflected in the query mapping table, as the training process can be completed in advance. In the application of the proposed method, the action output can be obtained only by querying the mapping table, and the the querying table complexity is $O(\log_2 N)$, where N is the number of elements in the mapping table. So the complexity of QLMTR is $O(\log_2 N)+O(n)$, and the complexity can be further reduced to $O(\log_2 N)$. Furthermore, since the state space of the proposed scheme is not large, the complexity is completely acceptable. The complexity comparison of the three methods is shown in Table 4.

Table 4. Complexity comparison of the three methods

	GPSR	D-S	QLMTR
complexity	$O(n)$	$O(n)$	$O(\log_2 N)$

It should be noted that, even in terms of complexity in application, the proposed method is more complex than the two comparison methods. Because the two comparison methods only need to obtain one or a limited number of parameters of the surrounding nodes without other calculation or storage, the complexity of the proposed method is still very low for the nodes and is completely acceptable.

Without loss of generality, this paper still uses Type III messages as an example to illustrate the advantages of algorithm (strategy) performance. **Fig. 9** shows the comparison between the proposed strategy and the GPSR algorithm in terms of transmission success rate and the comparison between the proposed strategy and the improved GPSR algorithm based on D-S evidence theory in terms of network overhead. As can be seen from the figure, the proposed strategy is significantly better than the GPSR algorithm which is sensitive to malicious nodes in terms of transmission success rate. At the same time, the strategy can guarantee the success rate of message transmission with relatively small overhead. As a whole, the proposed strategy could match with the type of transmission message, and can flexibly adjust and compromise between the transmission success rate and network overhead and other factors.

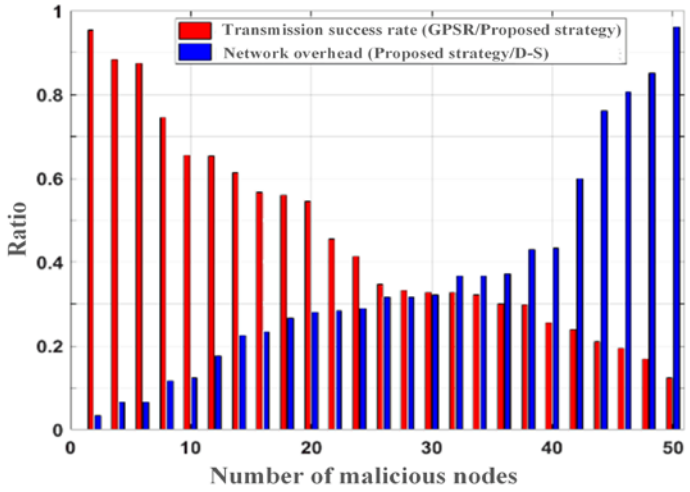

Fig. 9. Comparison of the proposed strategy and comparison method in terms of transmission success rate and network overhead

In order to facilitate the comparison of transmission delays, we choose a one-way message arrival process to calculate the time-consuming situation of transmitting messages. In order to eliminate the influence of the message transmission failure on the delay calculation process, we assume that the proposed strategy and the two comparison algorithms both adopt the method of sending unlimited times until the transmission is successful. Specifically, the transmission delay T_{delay} is defined as $T_{delay} = \left(\bar{N} - 1\right) \times T_{prepare} + \left(\bar{M} + 1\right) \times T_{action}$, where $\bar{N}$ is the average number of transmissions of the message, $T_{prepare}$ is the preparation delay between two repeated transmissions of the sending node, and $\bar{M}$ is the average number of forwarding nodes that participate in the process until the transmission is successful, T_{action} is the delay between transmission nodes related to the specific meta-action. See **Table 2** for the values of the above parameters.

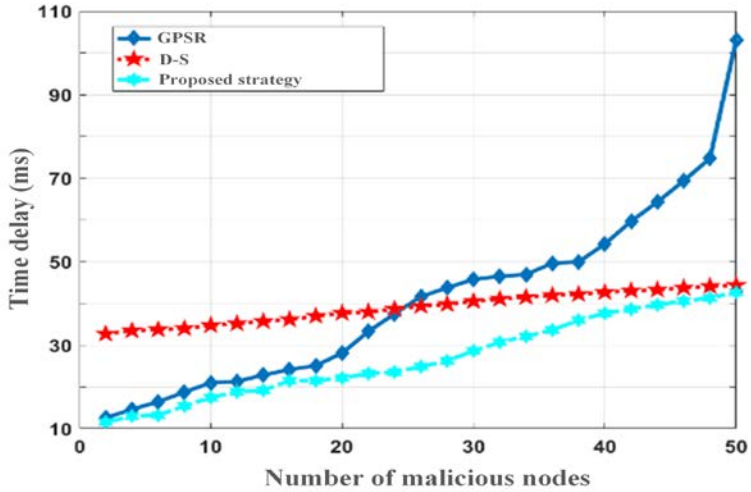

Fig. 10. Comparison of the performance of the three protocols in terms of transmission delay

Fig. 10 shows the comparison of the transmission delays for Type III messages between the strategy proposed in this paper and the two transmission protocols. It is not difficult to see from the figure, that with the increase of the number of malicious nodes, the message transmission delay of the three protocol strategies shows an obvious upward trend. Among them, the GPSR protocol is sensitive to malicious nodes, and its transmission times increase with the increase of malicious nodes, so the delay increases most rapidly. In addition, it can be seen from the figure that the strategy proposed in this paper combines the advantages of GPSR protocol with low transmission delay when the number of malicious nodes is low. With the increase of the number of malicious nodes, the proposed strategy can be used in the selection of the transmission protocol. The proposed strategy can tilt to the D-S meta-action in the selection of sending protocol, so that the increase in the number of malicious nodes will not show a rapid increase in time delay. It is not difficult to infer from the figure that when the number of malicious nodes further increases, the strategy proposed in this paper will abandon the GPSR meta-action, so its delay value will be consistent with the delay value of the improved GPSR algorithm based on the D-S evidence theory. Therefore, the strategy proposed in this paper is superior to the commonly used GPSR protocol algorithm and the improved GPSR algorithm based on D-S evidence theory proposed in this paper in terms of delay performance.

6. Concluding remarks

In reality, according to the characteristics of different message types, messages have different transmission requirements in VANET application scenarios. This paper considered the impact of malicious nodes on the security of the network, defined the message types. Meanwhile, by drawing on the characteristics of reinforcement learning that can spontaneously select actions to adapt to environmental needs through the exploration and utilization of the learning process, this paper has designed and proposed a set of message types as the state space. GPSR algorithm, improved GPSR algorithm based on D-S evidence theory and the combination of the two is the forwarding strategy of the action space. Numerous simulations have demonstrated that the proposed strategy can meet the spontaneous transmission requirements of different message types in the VANET scene. In addition, the strategy proposed in this paper can adjust the state space and meta-action elements according to the actual application, thus the strategy has broad application prospects.

Rererences

[1] Cyber Security Administration of the Ministry of Industry and Information Technology, White Paper on Network Security of Internet of Vehicles, 2020. [R/OL].
[2] M. Syfullah, J. M. Lim and F. L. Siaw, "Mobility-Based Clustering Algorithm for Multimedia Broadcasting over IEEE 802.11p-LTE-enabled VANET," *KSII Transactions on Internet and Information Systems*, vol. 13, no. 3, pp. 1213-1237, 2019. Article (CrossRef Link)
[3] A. BENGAG, M. E. Boukhari, “Enhancing GPSR routing protocol based on Velocity and Density for real-time urban scenario,” in *Proc. of 2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pp. 1-5, 2020. Article (CrossRef Link)
[4] R. A. Nazib, S. Moh, “Reinforcement Learning-Based Routing Protocols for Vehicular Ad Hoc Networks: A Comparative Survey,” *IEEE Access*, vol. 9, pp. 27552-27587, 2021. Article (CrossRef Link)
[5] Jianyong Li, *Research on optimization of learning algorithm for routing in wireless sensor networks*, Southwest University, 2016.

[6] Ming Yuan, *Research on Vanet Routing Algorithm Based on Reinforcement Learnin*, Xidian University, 2017.
[7] Celimuge Wu, Satoshi Ohzahata, and Toshihiko Kato, "Flexible, Portable, and Practicable Solution for Routing in VANETs: A Fuzzy Constraint Q-Learning Approach," *IEEE Transactions on Vehicular Technology*, 62(9), 4251-4263, 2013. Article (CrossRef Link)
[8] Celimuge Wu, Kazuya Kumekawa, and Toshihiko Kato, "Distributed Reinforcement Learning Approach for Vehicular Ad Hoc Networks," *IEICE Transactions on Communications*, E93.B(6), 1431-1442, 2010. Article (CrossRef Link)
[9] Plate R, Wakayama C, "Utilizing kinematics and selective sweeping in reinforcement learning-based routing algorithms for underwater networks," *Ad Hoc Networks*, 34(NOV.), 105-120, 2015. Article (CrossRef Link)
[10] G. Santhi, A. Nachiappan, M. Z. Ibrahime, R. Raghunadhane and M. K. Favas, "Q-learning based adaptive QoS routing protocol for MANETs," in *Proc. of 2011 International Conference on Recent Trends in Information Technology (ICRTIT)*, pp. 1233-1238, 2011. Article (CrossRef Link)
[11] Yiming Lin, *A Reinforcement Learning-based Routing Protocol in VANETs*, Xiamen University, 2017.
[12] Jinqiao Wu, Min Fang, Xiao Li, "Reinforcement Learning Based Mobility Adaptive Routing for Vehicular Ad-Hoc Networks," *Wireless Personal Communications*, 101(4), 2143-2171, 2018. Article (CrossRef Link)
[13] Shanshan Jiang, Zhitong Huang, Yuefeng Ji, "Adaptive UAV-Assisted Geographic Routing With Q-Learning in VANET," *IEEE Communications Letters*, vol. 25, no. 4, pp. 1358-1362, April 2021. Article (CrossRef Link)
[14] J. Aznar-Poveda, A. -J. Garcia-Sanchez, E. Egea-Lopez and J. Garcia-Haro, "MDPRP: A Q-Learning Approach for the Joint Control of Beaconing Rate and Transmission Power in VANETs," *IEEE Access*, vol. 9, pp. 10166-10178, 2021. Article (CrossRef Link)
[15] Network 5.0 Industry and Technology innovation Alliance, Network 5.0 Technology White Paper (2.0), 2021. [R/OL].
[16] C. Lin, S. Yuan, S. Chiu and M. Tsai, "ProgressFace: An Algorithm to Improve Routing Efficiency of GPSR-Like Routing Protocols in Wireless Ad Hoc Networks," *IEEE Transactions on Computers*, vol. 59, no. 6, pp. 822-834, June 2010. Article (CrossRef Link)
[17] Alsaqour, R., Abdelhaq, M., Saeed, R., Uddin, M., Alsukour, O., Al-Hubaishi, M. and Alahdal, T., "Dynamic packet beaconing for GPSR mobile ad hoc position-based routing protocol using fuzzy logic," *Journal of Network and Computer Applications*, vol. 47, pp. 32-46, 2015. Article (CrossRef Link)
[18] Wang, T., Anwar, S., Sun, H. and Zhou, Y., "Modified greedy perimeter stateless routing for vehicular ad hoc networking algorithm," *International Journal of Sensor Networks*, vol. 27(3), pp.163-171, 2018. Article (CrossRef Link)
[19] X. Yang, M. Li, Z. Qian and T. Di, "Improvement of GPSR Protocol in Vehicular Ad Hoc Network," *IEEE Access*, vol. 6, pp. 39515-39524, 2018. Article (CrossRef Link)
[20] H. Yuan, J. Geng, C. Liu, F. Bian, and T. Surapunt, "An improved GPSR routing algorithm based on vehicle trajectory mining," in *Proc. of the 5th International Conference Geo-Spatial Knowledge and Intelligence*, pp. 343–349, 2018. Article (CrossRef Link)
[21] C. Wang, Q. Fan, X. Chen, and W. Xu, "Prediction based greedy perimeter stateless routing protocol for vehicular self-organizing network," in *Proc. of IOP Conference Series: Materials Science and Engineering*, vol. 322, p.052019, 20108. December 2017. Article (CrossRef Link)
[22] A. Benmir, A. Korichi, A. Bourouis, M. Alreshoodi and L. Al-Jobouri, "An Enhanced GPSR Protocol for Vehicular Ad hoc Networks," in *Proc. of 2019 11th Computer Science and Electronic Engineering (CEEC)*, pp. 85-89, 2019. Article (CrossRef Link)
[23] S. Younes, M. Khelifi, A. Alioua and I. Souici, "EKF-GPSR: An Extended Kalman Filter for Efficient Routing in Vehicular Networks," in *Proc. of 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pp. 1-6, 2021. Article (CrossRef Link)

[24] M. Dabbaghjamanesh, A. Moeini and A. Kavousi-Fard, "Reinforcement Learning-Based Load Forecasting of Electric Vehicle Charging Station Using Q-Learning Technique," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4229-4237, June 2021.
Article (CrossRef Link)
[25] Zhihua Zhou, *Machine Learning*, Beijing: Tsinghua University Press, 2016.
[26] Yuanyuan Meng, Liancheng Min Xu, Ren, Yanfei Wang, "D-S Evidence Fusion Method Based on High Conflict Correction," *ComputerEngineering*, 44(1), 79-83, 90, 2018.
Article (CrossRef Link)
[27] K. Gu, X. Dong, X. Li and W. Jia, "Cluster-Based Malicious Node Detection for False Downstream Data in Fog Computing-Based VANETs," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 3, pp. 1245-1263, 1 May-June 2022. Article (CrossRef Link)
[28] A. M. El-Semary and H. Diab, "BP-AODV: Blackhole Protected AODV Routing Protocol for MANETs Based on Chaotic Map," *IEEE Access*, vol. 7, pp. 95197-95211, 2019.
Article (CrossRef Link)

Guoai Xu received his bachelor degree from Shangrao Normal University in June 1994, his master degree in Computational Mathematics from Huazhong University of Science and Technology in July 1998, and his doctor degree in information and Signal Processing from Beijing University of Posts and Telecommunications in August 2001.In September 2003, he was appointed as an associate professor of Beijing University of Posts and Telecommunications. In June 2008, he was appointed as a professor of Beijing University of Posts and Telecommunications.In the past 10 years, I have undertaken more than 30 research projects at provincial and ministerial level or above, including key RESEARCH and development projects of the 13th Five-Year Plan, and relevant achievements have been applied in typical users such as China Academy of Information and Communication Technology, CNCERT, China Mobile, China Construction Bank and so on. More than 60 patents have been applied in the past 5 years, 20 of which have been authorized. Participated in the formulation of more than 10 national standards and published more than 60 SCI retrieval papers. National Excellent Course award and Beijing Excellent Course award once each; Won the second prize of national teaching achievement once, the first prize of school teaching achievement twice and the second prize of school teaching achievement once. Supervised graduate students to receive several academic paper awards, including the Student International World Wide Web Conference Student Paper Award (the first in mainland China).

Boya Liu received the M.S. degree in electronics and communication engineering from Beijing Electronics Science &Technology Institute (BESTI), Beijing, China, in 2016. He is currently pursuing the Ph.D. degree in Cyberspace Security from Beijing University of Posts and Telecommunications (BUPT), Beijing, China. His research interests include wireless networking, VANET security and cryptography.

Guosheng Xu received the Ph.D. degree in Information Security from Beijing University of Posts and Telecommunications, China, in 2008. He is an IEEE member and a CCF member. Now, he is a senior lecturer and a master's supervisor in School of Cyberspace Security, Beijing University of Posts and Telecommunications. His current research interests include machine learning, intelligent software security and advanced cryptography.

Peiliang Zuo received the B.S. degree in information security from Xidian University, Xian, China, the M.S. degree in electronics and communication engineering from Beijing Electronics Science &Technology Institute (BESTI), Beijing, China, and the Ph.D. degree in information and communication engineering from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2013, 2016 and 2020 respectively. He is now a Lecturer at the department of Electronics and Communication Engineering, BESTI. His research interests include wireless networking, cognitive radio networks, network planning/optimizing, signal processing in wireless communications and software defined radio (SDR).