Reinforcement Learning Based Reliable Route Selection for Internet of Vehicles

Jung-Jae Kim, Minwoo Ryu, Si-Ho Cha

Abstract Self-driving cars have been receiving much attention recently, and communication problems between vehicles have become an issue. Due to frequent topology changes, routing problems occur in communication between vehicles. This is an important issue in the VANET (Vehicular Ad-hoc Network) and several papers have been presented to address this issue. However, existing papers are routing protocols that can resolve issues after they occur or only under certain circumstances, such as urban. Therefore, it is necessary to select the optimal relay nodes according to the circumstances surrounding the agent to ensure optimal performance at all times. For this purpose, this paper proposes RLSR (Reinforcement Learning based Selective Route Selection) algorithm that selects relay nodes through reinforcement learning. The algorithm proposed in this paper can ensure reliability by selecting the best relay nodes in any situation.

Keywords: Reinforcement Learning, VANET, Selective Routing, Adaptive Routing, Route Selection Routing

I. INTRODUCTION

The biggest topic in cars these days is self-driving cars that operate without human intervention. Self-driving cars focus on stability, convenience and efficiency in driving. The basic condition of self-driving cars is wireless communication between cars, which can quickly and quickly convey information about sudden increases in traffic, congestion, or accidents on the road. This can improve the stability and convenience of vehicle driving [1].The wireless network for data communication between these vehicles is referred to as the Vehicle Ad Hoc Network (VANET) [2]. VANET is a type of Mobile Ad home network (MANET), which consists of a number of wireless sensor networks, and forms a network between cars and road facilities to provide the necessary road and traffic information for driving the vehicle. It is also a service network that provides Internet access for internal users to use content. VANET's communication consists primarily of Vehicle to Vehicle (V2V) and Vehicle to Infrastructure (V2I), the communication between the vehicle and the road structure. The structure communicating with the vehicle is called the Road Side Unit (RSU), and the RSU is located in various locations such as intersections, highways, parking lots, and provides various information and services to the driver and passengers. This allows the driver of a vehicle to receive rapid increases in traffic, congestion, or accident information on the road[3].Routing algorithms are needed to send and receive data packets between cars [4]. The routing algorithm establishes a path from the sending node to the receiving node. Routing algorithms in VANET frequently change the network topology due to changes in vehicle speed and frequent movements. These topology changes are problematic, causing packet loss and delay, which reduces communication reliability and increases end-to-end delay. This makes geographic based routing protocols more suitable than conventional path-based routing protocols [5]. Geographic-based routing is a method of selecting the packet-delivery path based on the location information of the surrounding nodes and communicating the packet. Thus, there is no need to exchange separate control messages to maintain routes, and since there is no route navigation time, the overall communication speed increases as the topology changes quickly.The typical geographic-based routing method for VANET is Greedy Perimeter Non-Loading (GPSR) [6]. GPSR is basically sent in the greedy forwarding format (greedy mode). If the local maximal state is not allowed to perform the greedy forwarding because there is no forwarding node, stop the greedy forwarding and leave the local maximum through perimeter forwarding. The biggest barrier to routing through this GPSR is the local maximal. Local maximal raises the issue of link breaking, increasing packet Ross and packet delay. As a result, it interferes with the quick delivery of messages and reduces the overall transmission volume. Thus, previous studies tried to solve the local maximal by configuring the planar graph with the current vehicle situation in order to resolve the local maximal in GPSR. A study was also conducted to resolve the local maximal using a spanning tree that is easier to configure than a planner graph [7,8]. However, these methods have the overhead of constructing a graph or tree [9] and it also forwards the packet in the local maximal state until it approaches the intersection [10]. It uses digital maps to locate the intersection. These previously studied protocols require additional routing techniques, which results in high overhead. Also, with studies already operating after the local maximum is launched, it cannot be a fundamental solution to the local maximum. Therefore, a strategy is needed to avoid the occurrence of the local maximal as much as possible.In addition, information about adjacent

Revised Manuscript Received on May 22, 2019.

Jung-Jae Kim, Engineering Solution Office Control and Instrumentation Research Group,POSCO, South Korea

Minwoo Ryu, Service Laboratory Institute of Convergence Technology,KT R&D Center, South Korea

Si-Ho Cha,Dept. of Multimedia Science, Chungwoon University,Incheon, South Korea

* E-mail: shcha@chungwoon.ac.kr

nodes can be obtained through periodic non-context exchanges in the greedy mode. However, due to the time difference between the cycles of exchanging the non-contact messages, the node located outside the transmission is perceived as a nearby node due to the characteristics of the vehicle moving at fast speed. If the node selected as the relay node by greedy forwarding is a scale node, communication cannot continue. To address these problems, existing research uses different types of road (e.g. intersections, straight roads, highways, and highways) such as the method of selecting nodes for greedy mode at intersections [11] and urban centers. This selection of relay nodes via a heuristic approach does not always guarantee optimal performance. Therefore, it is necessary to react actively to the surrounding environment of the vehicle and to ensure performance by predicting the scale node.Therefore, in order to avoid the occurrence of local maximal as much as possible, this thesis proposes Refactoring Learning based on selection of relay nodes adaptively to the surrounding environment. The proposed protocol is based on Refactor Learning (RL), a type of learning method for machine learning, and constantly corrects or heuristic approaches to node selection. Therefore, vehicles with more than a certain weight learn gradually by strengthening learning to select the optimal number of nodes in each situation and select the optimal node among the surrounding nodes. The key idea of the proposed method is to adapt the road environment without human intervention. Thus, the ultimate objective of this paper is to propose an adaptive selective any-cast protocol in each road environment using RL which allows us to find the optimal number of nodes and to select the optimal relaying nodes in each environment. We summarize the major contribution of this paper as follows.

- Providing analytic approach through trial-and-error for VANET routing protocols.
- Proposing a RL-based intelligent adaptive broadcasting protocol for PBR as an alternative for the numerical method.
- Allowing for minimizing broadcast storm and fast delivering emergency message by using the optimal value, when the car changes the around environment at driving on the road.

The composition of this paper is as follows. Section 2 presents a brief overview of RLSR, and Section 3 describes how RLSR is designed for reinforcement learning and describes the RLSR algorithm. Section 4 describes the experimental results, and Section 5 describes the conclusions.

II. OVERVIEW

Reinforcement learning is a type of machine learning, in which an agent recognizes the current state of things in a certain environment and chooses the order of actions or actions that maximize rewards. Unlike supervised learning, which explicitly modifies behavior, Reinforcement learning focus on online training that enables continuous recognition of the environment and modification of one's own behavior. In other words, hardening studies learn in a Trial-and-error way, in which one's behavior is continually corrected by a reward after performing an action directly on the task. Thus, the protocols proposed in this paper use these characteristics of enhanced learning to select the optimal relay nodes based on the road conditions and the distribution of vehicles around the vehicle during routing of data packets.

2.1. RLSR Overview

The RLSR proposed in this paper uses a strategy to select the best node through learning in the environment to avoid local maximal by the scale node. Therefore, the RLSR proposed in this paper is based on nearby vehicle information and road information provided by the RSU to determine whether the relay node is the optimal node. Rewards given by reinforcement learning are based on turnaround time, which returns responses to the packet, to ensure speed and reliability at the same time. This improves network throughput by avoiding local maximal as much as possible by selecting the optimal relay node in a given environment. For this method, the RLSR proposed in this thesis consists of two phases: multipath preparation and node selection.

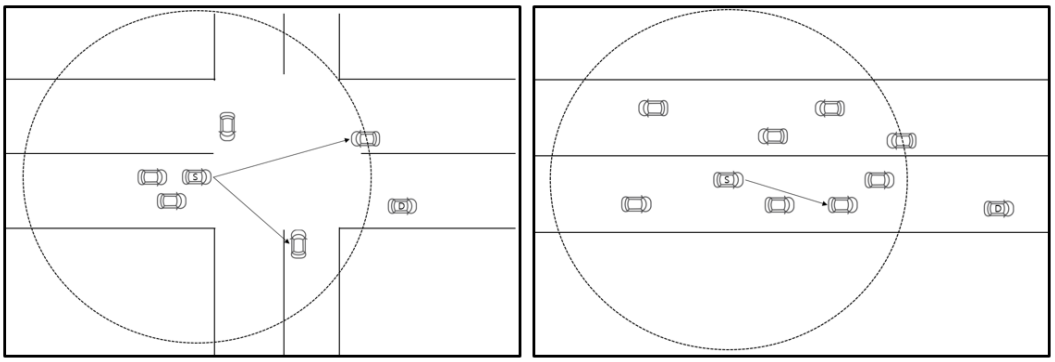

Figure 1. The Overview of RLSR

Figure 1 shows the basic routing process in RLSR. RLSR is a geo-based routing protocol using hardening learning. This requires two assumptions first. The first assumes that when there is a vehicle set V within the network with source vehicle s and target vehicle d, all vehicles know the position coordinates of their vehicle d with their neighbors. Second, the RSU near the road continuously transmits the shape of the road and its surroundings in a beacon message. The learning agent for source vehicle s is defined as the surrounding environment of the source vehicle s by the above assumption, the relative speed of each vehicle, the direction of the vehicle and the road shape.Route routing

in RLSR is based on the surrounding environment. When packets are to be delivered from the source vehicle s, the agent first checks the shape of the current road via a beacon message from the RSU. The agent then predicts how many nodes in the current environment can be selected through the number of nodes in the transmission, the shape of the road, the surrounding situation, and the speed of the agent to reach the target vehicle d. This is referred to as the multipath preset stage. The agent then determines whether the vehicle is the optimal node in the current situation by relative velocity of each vehicle inside the transmission range, direction of the vehicle, road shape, surrounding conditions, and the number selected through the multipath reception phase. This is called the node selection step. Each step describes the detailed procedures and methods through Section 4.1. After selecting a node, the agent receives compensation for selecting the node (behavior) and learns for further action. Therefore, Section 4.2 describes procedures and methods for setting up rewards for agents. In addition, the RLSR proposed in this paper uses two agents: multipath preparation phase and node selection phase, and the agents operate hierarchically. This is an overhead for communications that require real-time characteristics. Therefore, in this paper, two phases of learning are combined into one and behave similarly to multi agent referral learning. This has a huge advantage in learning speed by learning both agents at the same time as they are rewarded. This describes procedures and methods in Section 4.2.

III. REINFORCEMENT LEARNING DESIGN

Reinforcement learning depends on the definition of status and feedback in the environment where reinforcement learning is in. Therefore, a state and a compensatory function must be well defined in order to construct an efficient hardening learning algorithm.

3.1. State Design

The status space of the RLSR proposed in this paper can be divided into three types: the agent's own state and the surrounding environment, and the status of surrounding vehicles. First of all, the agent's own status includes the presence of abnormalities in the vehicle's internal drive and its speed and direction. And the environment around you means the situation in which the vehicle is in, so it includes how many vehicles there are capable of communicating, how many roads are shaped, and the situation around the road. Finally, the status of the surrounding vehicle includes the relative speed of the agent around the vehicle or the direction in which the vehicle is directed, and the reference direction in relation to the agent itself. The definition of state space used in RLSR is shown in table 1.

Table 1. The State Symbols of RLSR

Symbols	Definition	Note
SS_d	Agent internal drive unit malfunction	0 : Normal 1 : Strange
SS_s	One's own pace	
SS_a	One's own course	
ES_n	Number of vehicles within the communication range	
ES_r	A form of road	1:Highway 2 : Straight Road 3 : Cross Road 4 : Elevated Road
ES_e	Surroundings	1:Urban Area 2 : Open Area
VS_{rn}	n Relative velocity of the vehicle	
VS_{vn}	Direction of vehicle n	
VS_{an}	nThe reference direction for the vehicle	

3.2. Reward Design

Compensation given to the hardening learning structure through the final results is based on when an agent received a response packet for the transmitted packet within that local area. Therefore, vehicle agents within the network can be best rewarded by selecting the best agent within a continuous, interactive agent selection environment. This allows the vehicle agent to select the vehicle node that can deliver the packet most reliably in the current situation.The time that the *n*th vehicle V_n receives the response packet is equal to the current time minus the packet transmission time, expressed in equation (1).

$$V_n^t = cur_t - prv_t \quad (1)$$

However, in the case of such compensation, all vehicles within the network will be compensated on the same basis. However, packet transmission occurs over time. This means that although there is a significant time difference between the vehicle agent that initially sent the packet and the vehicle agent that exists at the end of the destination, it will be rewarded with the same value. Thus, in order to apply the time stream, the compensation of agents close to the destination must be different from the compensation of agents far from the destination. Thus, for this purpose, the agent's concept of subtracting the compensation it receives is added to enable each agent to receive a similar value reward. Thus, the compensation R_n^t of the *n*th vehicle V_n^t to which the discovery factor gamma has been added is as follows:

$$R_n^t = -|\gamma(V_n^t)| \quad (2)$$

This allows each agent to select its next relay node so that it has negative compensation for the relay node it chooses and the compensation it receives can be maximized (0).

3.3. Action Design

In RLSR, what each agent can do is determine the final selection of counts and medians to relay before the agent relays. By default, the number of agents is expressed as a vehicle that is located forward of the current node and is capable of transmitting packets. Also, the final selection for the relay can be expressed in two actions depending on whether the relay is to be made or not. Therefore, assuming that there are *n*

vehicles ahead of the k-node at the current point *t*, the actions that the k-node can perform A_t^k are as follows: Equation (3) is the behavioral space that the *k* node can take in the RLSR.

$$a_m = \begin{cases} 1 \\ \cdots \\ n \end{cases}$$

$$a_r = \begin{cases} 0 : no \\ 1 : yes \end{cases}$$

$$A_t^k = \{a_m, a_r\} \quad (3)$$

3.4. RLSR Algorithm

The RLSR consists of a structure that goes through two stages of multipath preparation and node selection to select the next relay node. The basic structure of the hardening learning model used in RLSR is shown in Figure 2.

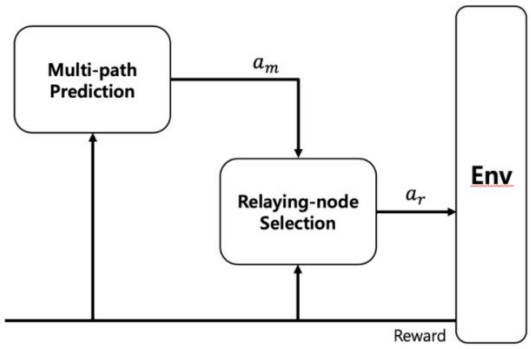

Figure 2. The Basic Structure of RLSR

Both steps use reinforcement to make decisions. In addition, learning is carried out using the Sarsa algorithm, which is the basic algorithm of enhanced learning. The following equation explains how to learn through the Sarsa algorithm.

$$S_m = \{SS_d, SS_s, SS_a, ES_n, ES_r, ES_e, VS_{rn}, VS_{vn}, VS_{an}\}$$

$$S_r = \{SS_d, SS_s, SS_a, ES_n, ES_r, ES_e, VS_{rn}, VS_{vn}, VS_{an}, a_m\}$$

$$S = \begin{cases} S_m \\ S_r \end{cases}$$

$$V(S) = V(S) + \alpha[R + \gamma V(S') - V(S)]$$

$$S = S' \quad (4)$$

The above process continuously updates the value of any behavior (number of nodes, selection of relay nodes) in the current state by the vehicle agent. This serves to reinforce one's behavior by continuing to act in the current state if the compensation currently obtained is high. The table 2 shows the RLSR algorithm.

Table 2. The RLSR Algorithm

1.	Sensing of Surrounding Environments
2.	$a_m \leftarrow$ action given by S_m
3.	flag = 0
4.	For j = 1 to n
5.	$S_r = \{SS_d, SS_s, SS_a, ES_n, ES_r, ES_e, VS_{rj}, VS_{vj}, VS_{aj}, a_m\}$
6.	Take action a_r by S_n
7.	If a_r == 1 then
8.	flag += 1
9.	If flag == a_m then
10.	Break
11.	End If
12.	End If
13.	End For
14.	observe reward, *R*, and next state, S_m'; next state, S_r'
15.	$V(S_m) = V(S_m) + \alpha[R + \gamma V(S_m') - V(S_m)]$
16.	$S_m = S_m'$
17.	$V(S_r) = V(S_r) + \alpha[R + \gamma V(S_r') - V(S_r)]$
18.	$S_r = S_r'$

Line 1 indicates that the packet is sent firstly sensitive to the surrounding environment. The second line determines how many nodes a packet should be sent in the current situation based on its sensitive state. Line 3 is a flag to select as many as the number of relay nodes currently determined. Paragraphs 4 through 13 describe the part of making decisions about whether to make a relay node. Line 5 constructs for the nearest node based on the number of relay nodes determined in line 2. In line 6, the decision is made on whether to use to include the node in the relay node. Paragraphs 7 to 12 describe the stopping of decision-making if the number of transit node candidates was chosen as the number of relay nodes determined in . Paragraphs 14 to 18 describe the portion of weight update by Sarsa algorithm of reinforcement learning.

IV. SIMULATION AND EXPERIMENT

Experimental results show that GPSR and RLSR. GPSR is representative geographical routing protocol. And they compared with straight roads and intersections. There are 30 vehicles on a road of about 1000 meters, and the vehicle moves at a speed of 50 to 100 km/h. at this time, the packet transmission speed on the straight road and the intersection is compared. However, RLSR is a state in which learning is completed. Table 3 shows the simulation results.

Table 3.Experiment results of RLSR and GPSR

Protocol		Minimum transmission time	Maximum transmission time	Average transmission time
RLSR	Straight	0.6	0.8	0.692
	Intersection	0.6	0.9	0.748
GPSR	Straight	0.6	1.0	0.874
	Intersection	0.6	1.3	1.144

RLSR and GPSR are the same for best case (Minimum transmission time). In the case of GPSR, the local maximum occurs at the intersection, which shows a very large difference of time. On the other hand, in the case of the RLAR, the relay node is selected based on the knowledge of the direction and the direction of the vehicle in consideration of the speed and direction of the surrounding vehicle. And we measure the average time, we can see that RLSR is faster than GPSR. However, as a result of the above experiment, the agent tends to select a large number of nodes in front as learning progresses ($a_m \approx maximum$). This means that the more packets are delivered to many nodes in the simulation, the more reliable the packets reach the destination. However, in a real environment this is the same as full broadcast, which causes a broadcast storm.

V. CONCLUSION

The proposed RLSR is an algorithm for solving the problem caused by other protocol depending on the local maximum avoidance problem and the existing routing protocol. Thus, it is possible to secure the reliability of the routing path by selecting the relay node according to the surrounding environment. We also verify this through simulation. However, the RLSR protocol proposed in this paper has a problem that the number of relay nodes is increased as learning progresses. If all the relay nodes transmit in the algorithm, it becomes meaningless because it becomes a full broadcasting routing protocol. Therefore, in future research, it is necessary to consider the state of the surrounding network in decision making by adding the current communication state to the state space in order to solve the problem.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2016R1D1A1A09917662).

REFERENCES

1. Brian Paden, Michal Čáp, Sze Zheng Yong, Dmitry Yershov, Emilio Frazzoli. A Survey of Motion Planning and Control Techniques for Self-driving Urban Vehicles. 2016 Jun;1(1):33-35.
2. Hannes Hartenstein, Kenneth Laberteaux. VANET: Vehicular Applications and Inter-Networking Technologies:Chichester, U.K.: Wiley; c2010. p. 24.
3. Qutaiba I. Ali. Security Issues of Solar Energy Harvesting Road Side Unit(RSU). Iraq J. Electrical and Electronic Engineering. 2015;11(1):18-31.
4. Yongmei Sun, ShuyunLuo, Qijin Dai, YuefengJi. An Adaptive Routing Protocol Based on QoS and Vehicular. 2015 Jun;2015:13.
5. SeokYoon Kang, JeongCheol Lee, Ki-Il Kim, HoSung Park. Region-Based Collision Avoidance Beaconless Geographic Routing Protocol in Wireless Sensor Networks. 2015;15(6):13222-13241.
6. Alroqi, Y. M. A Novel Ferry Assisted Greedy Perimeter Stateless Routing protocol (FA-GPSR) for Ad-hoc Networks in Remote Locations [PhD's thesis]. [Nottingham (PA)]: Nottingham Trent University; 2015 Apr p.184.
7. Byunggeon Lee, Wonsik Yoon. Power Aware Greedy Perimeter Stateless Routing Protocol for Wireless Ad Hoc Network. The Institute of Electronics Engineers of Korea – Telecommunications. 2008;45(7);62-66.
8. Peppino Fazio, CesareSottile, Amilcare Francesco Santamaria, Mauro Tropea. Vehicular Networking Enhancement and Multi-Channel Routing Optimization, Based on Multi-Objective Metric and Minimum Spanning Tree. Advances in Electrical and Electronic Engineering. 2013;11(5):349-356.
9. Jong-Hyun Kim. Kee-Cheon Kim. Woo-Young Jung. Grid-based Location Service Spot Scheme for Optimized Routing Path on VANET. The Korea Institute of Intelligent Transport Systems. 2010 Feb;9(1):76-90.
10. VoichiţaRoib, IlincaRoib. Contributions Regarding the Creation of the Digital Map of the Public Transport in the Metropolitan Area of Cluj. Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca: Horticulture. 2018;75(1):70-72.
11. Do-SikAn, Gi-Hwan Cho. Message Delivery Schemes in Robust of Network Disconnection for GBSR based VANET Routing. The journal of multimedia information system. 2009 Nov;137-140.