Reinforcement Learning based Gateway Selection in VANETs

Original Scientific Paper

Hasanain Alabbas
Budapest University of Technology and Economics,
Faculty of Electrical Engineering and Informatics, Department of Networked Systems and Services
H-1117, Budapest, Hungary
Al-Qasim Green University
Computer Center
Babel, Iraq
hasanain@hit.bme.hu

Árpád Huszák
Budapest University of Technology and Economics,
Faculty of Electrical Engineering and Informatics, Department of Networked Systems and Services
H-1117, Budapest, Hungary
huszak@hit.bme.hu

Abstract *– In vehicular ad hoc networks (VANETs), providing the Internet has become an urgent necessity, where mobile gateways are used to ensure network connection to all customer vehicles in the network. However, the highly dynamic topology and bandwidth limitations of the network represent a significant issue in the gateway selection process. Two objectives are defined to overcome these challenges. The first objective aims to maximize the number of vehicles connected to the Internet by finding a suitable gateway for them depending on the connection lifetime. The second objective seeks to minimize the number of connected vehicles to the same gateway to overcome the limitation of gateways' bandwidth and distribute the load in the network. For this purpose, A gateway discovery system assisted by the vehicular cloud is implemented to find a fair trade-off between the two conflicting objectives. Proximal Policy Optimization, a well-known reinforcement learning strategy, is used to define and train the agent. The trained agent was evaluated and compared with other multi-objective optimization methods under different conditions. The obtained results show that the proposed algorithm has better performance in terms of the number of connected vehicles, load distribution over the mobile gateways, link connectivity duration, and execution time.*

Keywords*: Gateway selection, Reinforcement learning, Proximal policy optimization, VANET*

1. INTRODUCTION

1. INTRODUCTION

Vehicular Ad hoc network (VANET) is one of the interesting fields in the Intelligent Transportation System (ITS) that exploits the moving vehicles as mobile nodes in the network. Because of its wide applications of increasing safety for drivers, reducing car accidents, and providing Internet to users, it has attracted researchers' interest [1]. VANET infrastructure is composed of two communication entities, On-Board Unit (OBU), which is integrated inside the vehicles, and Road Side Unit (RSU), which is mounted on the roadsides or near traffic intersections [2]. These communication entities allow two types of communications. Vehicle-to-Vehicle (V2V), which enables the vehicular nodes to contact each other directly, and Vehicle-to-Infrastructure (V2I), in which vehicles can communicate with RSUs [3]. Providing vehicles with a permanent internet connection has become an urgent necessity to feed drivers with relevant road information and offer a comfortable trip for passengers [4][5]. Providing the Internet for vehicles requires finding a suitable gateway. Unfortunately, the implementation of this goal is facing many challenges, most notably the highly dynamic topology of the network and the bandwidth limitations [3][6][7]. Most gateway discovery techniques are based on Inquiry and Solicitation messages sent and received between the vehicular nodes to find a suitable gateway [8][9]. These techniques have many issues (broadcast storm problem, overhead) when the nodes increase [10]. The progress in cloud computing and making it compatible with ITS provides a valuable opportunity to benefit from cloud computing resources utilized by VANET services [11]. Many research efforts have been made in this area, which produced a new paradigm called Vehicular Cloud (VC) [12][13]. VC offers many features such as collecting vehicle information, optimizing traffic control, and detecting congestion [14]. Due to the massive services and features provided by VC, some studies have invested it to perform more complex computations and find efficient solutions to improve gateway selection and address overload problems. However, the disadvantage of these solutions is that they do not find a gateway with the highest link connectivity duration (LCD), because they don't take the nature of roads and their vulnerability to traffic congestion into account. The efficiency of these solutions decreases in urban area so that the execution time increases catastrophically with the number of vehicles. VC is exploited to build a novel model by using reinforcement learning. The proposed model is designed to optimize gateway discovery by maximizing LCD and minimizing bandwidth overload. For this purpose, Multi-Objective Reinforcement Learning (MORL) is used. In this paper, the Proximal Policy Optimization (PPO) model is adopted to train the agent due to its better performance compared to other standard reinforcement learning algorithms. To the best of our knowledge, the previous studies of gateway selection use current speed, direction, and distance as crucial factors in the selection, ignoring other factors that have a significant influence on selection like road density and intersections. The main contributions of this paper are as follows:

1. The algorithm implicitly takes into consideration factors related to road density and the impact of intersections in addition to the traditional factors (speed, distance, and direction).
2. The decision of electing the gateways is based on the real and actual time of the link connectivity duration between the vehicles, so the algorithm gives better results in terms of stability and scalability.

The rest of this paper is arranged as follows. Section 2 reviews some related literature on VANET and gateway selection solutions. Section 3 presents the proposed system model components in detail and discusses the reinforcement learning method used in the model. Section 4 evaluates the presented technique with other existing multi-objective optimization solutions. The conclusion and future work are presented in section 5.

2. RELATED WORKS

Providing permanent access requires finding a suitable gateway, which has a direct connection to the Internet. In literature, the gateway can be a stationary station (RSUs, cellular base stations), which is considered as a part of vehicular network infrastructure [15]. In [16], the authors suggested a Fuzzy QoS-balancing Gateway Selection algorithm to connect the vehicles to the LTE infrastructure. The proposed algorithm employs the distributed LTE Advanced eNodeBs as stationary gateways to meet the vehicles' needs. The connections between the vehicles and LTE advanced eNodeBs are either directly or by choosing a relay gateway. Fuzzy logic is applied to select the best gateway based on signal strength, load, link connectivity duration, and QoS traffic classes. However, the drawback of these kinds of solutions is the handover of connections that are generated as a result of the vehicles' high speed compared to the fixed road infrastructure. Moreover, the gateway selection process uses a reactive approach, where the vehicles broadcast the Solicitation message to seek a suitable gateway, which engenders a high amount of overhead.

The study [17] proposed a routing protocol for mobile gateway discovery to ensure Internet access for vehicles in the area where it is not available. Many parameters have been adopted, namely robust parameters, like received signal strength (RSS), trust connection, the number of hops, and route lifetime, to establish a robust route protocol for mobile gateway discovery. Two stages are defined to set the routing protocol. The first stage is the gateway selection process which starts when the moving vehicles toward the UMTS base station enter its coverage area. These vehicles declare themselves as mobile gateways if the received signal strength of UMTS is greater than the RSS threshold. Relays selection represents the second stage in which the mobile gateways select relays based on robust parameters. The simulation results exhibit good performance in terms of packet delivery and decreasing the overhead when applied in a highway scenario. However, the proactive and reactive strategies used can decrease the throughput when the vehicle's number increases. Idrissi et al. [18] used the vehicular cloud architecture to develop the gateway discovery system. They adapted a multi-criteria decision approach known as Preference Ranking Organization METHod for Enrichment of Evaluations (PROMETHEE) to find the best gateway. Several criteria are considered, representing the difference between the customer vehicles that tend to get access to the Internet and the mobile gateways that have a direct Internet connection. These criteria have been used to increase the number of connected customer vehicles and decrease the traffic routed by the mobile gateway. However, the gateway selection mechanism in this study lacks the use of optimization techniques. Sara Retal and Abdellah Idrissi proposed a method to improve the mobile gateways selection [19]. Multi-Objective Optimization is considered to overcome the weakness of the previous study by maximizing the number of connected vehicles and minimizing the overload of the gateways. Different models are used to find the best solution, which represents a trade-off solution of different conflicting objectives. To the best of our knowledge, this study is considered one of the pioneering studies in the scope of mobile gateway selection techniques; therefore, we will adopt one of its used model, namely Integer Optimization Problem (IOP), as a benchmark. This algorithm can perform well when vehicles keep a relatively constant speed and direction, especially on highways. Still, they do not perform well in urban areas because of the roads' nature and intersections, which significantly affect vehicles' variation of speed and direction. Moreover, the execution time of the gateway selection process is relatively high, and it increases significantly when the vehicles increase. A novel model is presented to discover the mobile gateways by using reinforcement learning.

3. GATEWAY SELECTION ARCHITECTURE

The main objective of this study is to find a suitable Mobile Gateway (MG) upon request from Client Vehicles (CVs). MGs are vehicles with direct Internet access, whereas CVs represent all vehicles that have no direct connection. In the analyzed urban scenario, we assume that the public transport buses are equipped with Internet access and can serve as MGs. Their convenient speed, which usually does not exceed 40 km/h, and their regular geographic distribution in urban areas make them have a highly predictable day-to-day pattern [20]. As shown in Fig. 1, the proposed system consists of CVs, MGs, VANETs infrastructure (4G/5G base station, RSU), and Vehicular Cloud (VC). The CVs and MGs can communicate with each other via V2V connection, while VANETs infrastructure represents the link between the VC from one side and CVs and MGs on the other side. VC consists of two servers in which, The Registrar server monitors the VANETs environment, collects the vehicles' information, and registers it in a dataset. In contrast, the Agent server is in charge of gateway discovery for vehicles trying to access the Internet.

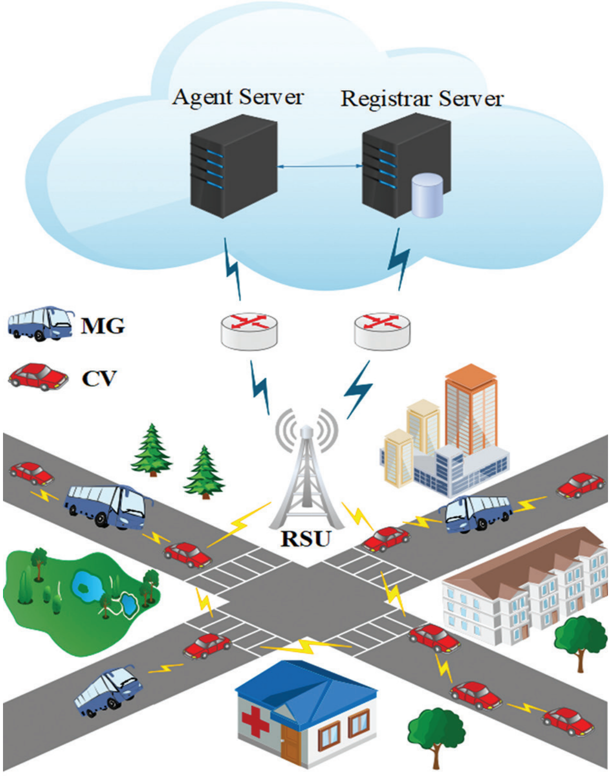

Fig. 1. System architecture.

The Registrar Server collects the necessary information on CVs and GWs related to speed, geographical location, direction, and the link connectivity duration (LCD) between CVs with all MGs. This information is stored in a database containing records generated for each CV. Each registry contains the difference between one CV and all the MGs in terms of the geographical location (longitude and latitude), speed, direction, and the traffic amount routed by the MG and LCD. This information is collected and stored in the database periodically so that as soon as the process of filling in the data of the current record is completed, the process of adding a new record begins. The gateway discovery system is built in the Agent server by using reinforcement learning. The main goal was to achieve two contradicting objectives by finding the best trade-off between them. These objectives are:

- Objective 1: Increasing the number of CVs connected to the MGs with the highest LCD.
- Objective 2: Minimizing the traffic volume routed by MGs by decreasing the number of CVs connected to the same MG.

In the training phase, the RL agent starts to adapt and learn from the environment of VANET based on the data collected by the registrar server. The RL agent will be able to find the best MG for each CV when the training phase ends. The role of the Agent server is to select the best GW for each CV requesting Internet access.

3.1. REINFORCEMENT LEARNING

Reinforcement learning (RL) is a branch of machine learning that imitates human behavior in acquiring skills by planning for the future and deciding based on it in a specified environment. The main objects in an RL problem are the agent and the environment. The concepts (state (*S*), action (*A*), reward (*R*)) represent the interaction of the agent with its environment. The agent monitors the environment state (s_t) and takes action (a_t) at the time (*t*), which causes a state transition to a new state (s_{t+1}). The correctness of the decision taken is determined by the reward (rt) given to the agent. The reward function *R*(*r*|*s, a, s'*) represents the immediate reward probability for state transition [21], [22] as shown in Equation (1):

$$R(r \mid s, a, s') = \Pr\left(r_t = r \mid s_t = s, a_t = a, s_{t+1} = s'\right), \quad (1)$$

the policy *π(a, s)* defines the behavior of the agent depending on its observations. The mapping between the action (*a*) and the state (*s*) is modeled by the policy *π(a, s)*, which represents the action (*a*) probability as follows:

$$\pi(a \mid s) = \Pr\left(a_t = a \mid s_t = s\right), \quad (2)$$

the agent explores the optimal policy *π*(a, s)* by maximizing accumulated discounted reward for each *s* ∈ *S* and *a* ∈ *A* shown in Equation (3):

$$\pi^*(a \mid s) = \underset{\pi(a|s)}{\arg max} \sum_{t=t_0}^{t_{end}} \gamma^{t-t_0} r_t, \quad (3)$$

where $\gamma \in (0,1)$ is the discount factor and t is the time horizon. Policy optimization algorithms can be classified into two categories which are value-based algorithms and policy-based algorithms. Compared with the value-based algorithm, policy-based algorithms have better convergence and are more convenient for large action spaces. Proximal Policy Optimization (PPO) [23] algorithm is an actor-critic method that combines the value-based and the policy-based algorithm. Two neural networks are applied. The first one, named *actor*, takes the state *(s)* as entries and outputs the policy *π (a, s)*, while the second one, named *critic*, optimizes *V(s)* that measures the goodness of the action (a). PPO uses the advantage *A(s, a)* to reduce the estimation variance, which is expressed in the following:

$$A(s_t, a_t) = Q(s_t, a_t) - V(s_t) \ , \tag{4}$$

$$Q(s_t, a_t) = r_t + \sum_{i=1}^{T-1} \gamma^i r_{t+i} + \gamma^{t+T} V(s_{t+T}) \ , \tag{5}$$

where *Q(s,a)* represents the cumulative discount reward when action at is taken for the state s_t, while *V(s)* represents the baseline, this technique allows updating the policy network in a direction that chooses better actions. PPO uses the trust-region (TRPO) method to ensure that the new updated policy never goes far away from the current policy, making it more stable and reliable. The primary objective function of PPO is denoted as $L^{CLIP(\theta)}$:

$$L^{\text{CLIP}(\theta)} = \mathbb{E}_\tau\left[min\left(R_t A_t, \text{clip}\ (R_t, 1-\epsilon, 1+\epsilon)\right) A_t\right], \tag{6}$$

where A_t is an abbreviation of $A(s_t, a_t)$, ϵ denotes a small positive constant, and the policy ratio (R_t) measures the similarity between the updated policy and old policy as shown in Equation (7):

$$R_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \ , \tag{7}$$

while clipping function, clip (R_t, *1-ϵ, 1+ϵ*) ensures the R_t moving inside the interval [*1- ϵ, 1+ ϵ*]. For these reasons, the PPO algorithm is adopted in the proposed gateway selection system, namely PPO-GS. Three components should be defined carefully to enable the agent to sense the environment and make the right decision: state, action, and reward.

3.2. DEFINITION OF OBSERVATION STATE

The vehicles in VANET are classified into two categories MGs and CVs. The state (st) will be created for each CVi that needs Internet access and looks for a connection to a suitable MGj. It represents the relationship between the CV and all the MGs in terms of geographical location, speed, and available bandwidth. The state is expressed by the entries as follow:

$$X = \begin{bmatrix} Lo_{i1} & Lat_{i1} & V_{i1} & \theta_{i1} & T_1 \\ Lo_{i2} & Lat_{i2} & V_{i2} & \theta_{i2} & T_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Lo_{ij} & Lat_{ij} & V_{ij} & \theta_{ij} & T_j \end{bmatrix} \tag{8}$$

- $Lo_{ij} = Lo_i - Lo_j$, where Lo_i and Lo_j denote the longitude of CV_i and MG_j, respectively.
- $Lat_{ij} = Lat_i - Lat_j$, where Lat_i and Lat_j denote the latitude of CV_i and MG_j, respectively.
- $V_{ij} = V_i - V_j$, where V_i and V_j denote the velocity of CV_i and MG_j, respectively.
- $\theta_{ij} = \theta_i - \theta_j$, where θ_i and θ_j denote the direction of CV_i and MG_j, respectively.
- T_j denotes the traffic volume routed by MG_j.

The difference in longitude and latitude is applied between CVs and MGs rather than distance. By using the difference of the coordinates, the algorithm can decide whether the MG is at the front of the CV or not. The number of entries in the state increases with the increase in the number of MGs since the state represents the relationship of each CV to all MGs. Since the relationship of the CV to each MG is represented by five parameters *S*=(Lo,Lat,V,θ,T), the total number of entries to represent the state is |*S*|·|*MG*|,where |*S*| is the number of parameters used to describe a state, while |*MG*| is the number of MGs. The number of GWs distributed in the environment is 20, so the agent state is composed of 100 entries.

3.3. DEFINITION OF AGENT ACTION AND REWARDS

The decision taken in the gateway selection system is defined as the agent actions. In the proposed system, the set of actions represents the number of MGs. Action $a = \{a_1\ a_2\ a_3 \dots a_n\}$, where a_1 represents the selection of MG_1 while a_n represents the selection of MG_n. The reward is assigned based on two metrics: the first metric is the connection lifetime between the CV and MG, whereas the second is the traffic amount routed by each MG. The first parameter aims to increase the number of vehicles connected to the Internet, while the second one aims to reduce the CV_s linked to the same MG. Setting the reward function with two contradicting objectives needs to apply Multi-Objective Reinforcement Learning (MORL). MORL aims to find a trade-off solution for multiple conflicting objectives. Several approaches can be used to achieve this goal [24]. Weighted Sum Approach (WSA) is applied to solve the multi-objective problem in the reward function. The weight (*w*) of each metric is set to define the objectives preferences as shown in Equation (9):

$$R = w_1 \cdot lcd_{ij} + w_2 \cdot T_j, \tag{9}$$

where lcd_{ij} denotes the link connectivity duration value between CV_i and MG_j, while T_j represents the traffic volume routed by MG_j. w_1 and w_2 take values between 0 and 1 depending on the importance of the objectives so that $w_1 + w_2 = 1$.

The reward value is positive when the action is valid. Else, the reward is negative. The positive reward ranges in value between 0 and 20, while the negative reward is (-4).

The negative reward is applied in two cases:

1. The agent selects an MG which is out of the CV coverage area.
2. The agent selects an MG that does not have enough amount of traffic.

Setting the reward in this manner motivates the agent to find an MG with the best link connectivity duration and the least amount of traffic handled by it. In contrast, the negative reward ensures that the agent avoids choosing an MG out of the communication range of CV, or an MG cannot dedicate a channel to a CV. As is known, there are no rules that specify the reward value. Any value can be used, provided the agent is learning correctly. The reward value was adopted as mentioned above after training the agent several times with different rewards values because the assumed value showed a faster response from the agent to learning.

3.4. AGENT STATE PARAMETERS

In the proposed system, training the agent relies entirely on the dataset generated by the registrar server. The dataset represents an enormous number of snapshots taken from the VANET environment. Each entry in the dataset consists of two main parts: 1) the first part represents the state s, which involves the amount of difference for each CV with all MGs in terms of longitude, latitude, speed, and direction, as well as the available amount of bandwidth for each MG. 2) the second part which contains the LCD values. PPO is employed to maximize the MGs selection return. The reward r is a multi-objective reward in which the agent tends to find an MG for a CV with the maximum LCD and minimum number of CVs connected to it. The episodic environment is considered during agent training so that each episode consists of 100 steps (finding MGs for 100 CVs). The VANET environment is variable and highly dynamic as it depends on moving nodes in which it is difficult to determine the future return. So in an environment as dynamic as the presented work, it is appropriate for the RL agent to maximize the current reward rather than the cumulative discount reward. This is done by adopting $\gamma = 0$. To prove our hypothesis, we applied different gamma values during the training process. We found that the lower the gamma value, the faster the agent learns. Fig. 2 shows the highest reward collected when $\gamma = 0$.

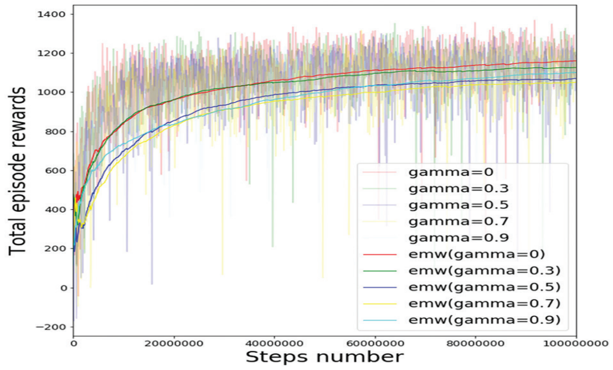

Fig. 2. Training the agent with different gamma values

4. RESULTS

This section exhibits the performance evaluation of the proposed algorithm. For comparison purposes, the well-known gateway selection method (A multi-objective optimization system for mobile gateways selection in vehicular Ad-Hoc networks) [19] is simulated, namely the MOO algorithm. The MOO technique has similar properties as this work. Therefore, it is adopted as a benchmark. It has a centralized algorithm to make a decision, while the rest of the recent studies are decentralized algorithms in which the gateway selection depends on sending messages between vehicles. This work is implemented by using the Python programming language. The stable-Baseline3 library is used to implement and train the RL agent [25], whereas Gurobi Optimizer is executed to solve the multi-objective optimization problems used in the MOO solution. The vehicles' mobility and their behavior are simulated by using SUMO. The simulation is performed under an urban area map of 1500 m x 1500 m using Open Street Map (OSM). OSM provides free maps from all around the world, which makes the simulation more realistic. The urban area environment is adopted as shown in Fig. 3 because it can be taken as a true scale of how successful an algorithm is. It is more challenging than the highway environment because of the road's nature (congestion in a rush, speed limit, intersections, etc.).

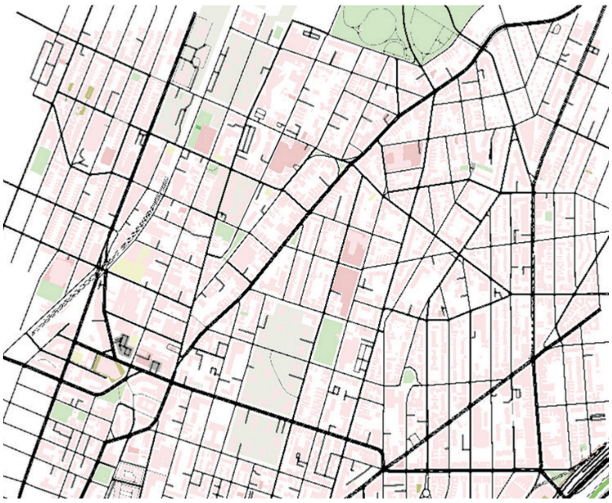

Fig. 3. Map from OSM

The number of MGs deployed in the simulation network is 20, while the number of CVs is 60-100. They are deployed randomly in the simulation to evaluate all the proposed models and compare them with the MOO method. By fixing the number of MGs, whereas the CVs number is variable. The entire simulation parameters are listed in Table 1. The proposed models are compared to the MOO models (MOO1, MOO2, and MOO3) to evaluate the performance. Several metrics have been used for the performance evaluation, including the number of connected vehicles (CVs), the distribution of CVs among MGs, LCD between CVs and MGs, and the execution time.

Table 1. Simulation Parameters

Parameters	Setting
Transmission Range	500 m
X-coordinate	0-1500 m
Y-coordinate	0-1500 m
Vehicles speed	0-20 m/s
MGs Number	20
CVs Number	60-100

In the simulation, different weights are applied, as mentioned in section 3.3, to determine the objectives preferences by utilizing the weighted sum approach. According to these weights, three methods are implemented in the presented study: PPO-GS1, PPO-GS2, and PPO-GS3. In PPO-GS1, the reward function maximizes the first objective only. In PPO-GS2, the reward function takes into consideration the two objectives but with more preference for the first one. Meanwhile, the objectives in PPO-GS3 take the same priority. On the other hand, three approaches which are called MOO1, MOO2, and MOO3, are defined in the MOO algorithm as detailed in Table 2.

Table 2. Objectives weights

		W1	W2
PPO-GS1	MOO1	1	0
PPO-GS2	MOO2	0.7	0.3
PPO-GS3	MOO3	0.5	0.5

PPO-GS1 has the highest number of connected CVs because it makes the decision based on objective1 and does not consider the MGs bandwidth limitation. The relying on objective1 in selecting MGs causes inequality and a wide variation in the distribution of CVs over the MGs, as shown in Fig. 4.

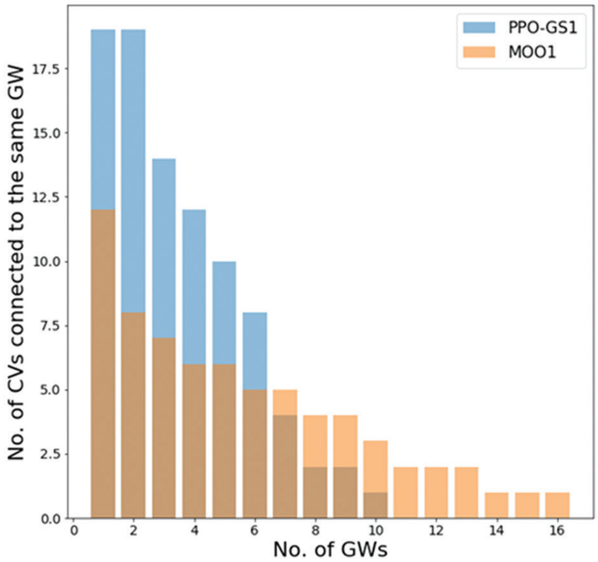

Fig. 4. CVs distribution with w_1=1 and w_2 =0

Fig. 5 and Fig. 6 show that the number of CVs connected to the same MGs decreases as the weight of the second objective increases, which takes into consideration the limitation of MGs bandwidth; therefore, PPO-GS3 and MOO3 exhibit more equitable distribution in comparison with other solutions.

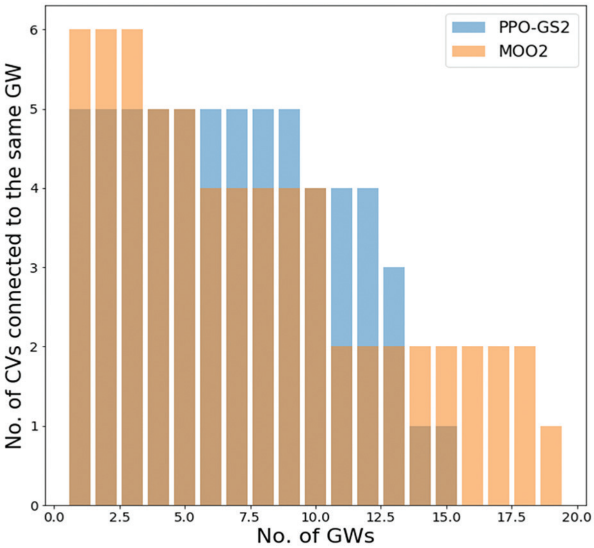

Fig. 5. CVs distribution with w_1=0.7 and w_2 =0.3

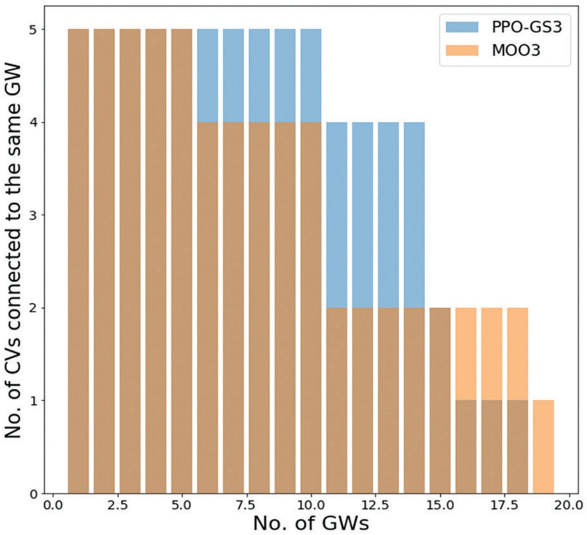

Fig. 6. CVs distribution with w_1=0.5 and w_2 =0.5

Fig. 7 shows that the new proposed technique, in general, has better performance in increasing the number of connected CVs in comparison with MOO solutions. The reason is due to the fact that the MOO algorithm uses the constraints to ensure that the MGs and associated CVs have similar speeds and directions. Consequently, CVs with a large difference in speed and direction cannot find a suitable MG. All the approaches were tested applying the same conditions. Each scenario was executed and evaluated multiple times so that each point in the plot shows the mean of 15 executions. The number of CVs in each scenario is varied from 50 to 100, whereas the number of MGs is fixed to

20. Most of the charts are submitted with a 95% confidence interval. Regarding LCD, it is essential to find MGs to the CVs with the highest LCD because this ensures stable Internet connections for CVs.

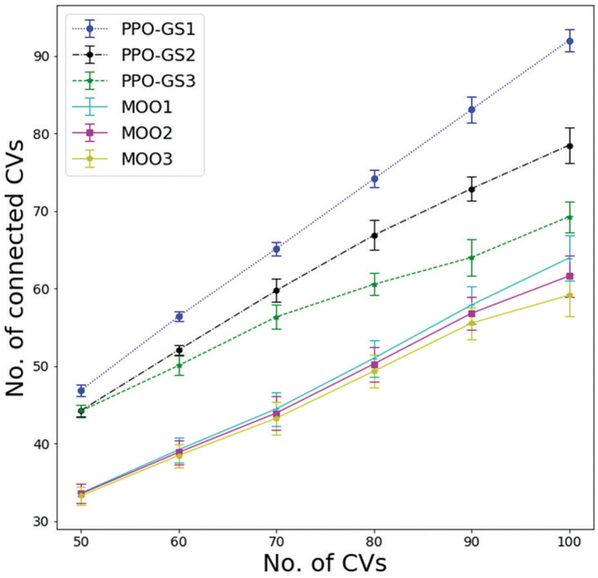

Fig. 7. Number of connected CVs

As Fig. 8 shows, the percentage of MGs selection with the best LCD in the new proposed algorithm is not affected by the increase of the number of CVs compared to the MOO algorithm, which decreases when the number of CVs increases. In Fig. 9, the execution time of the proposed algorithm is low and almost unaffected by increasing the CVs number. In contrast, the MOO solution's execution time is high and affected drastically by increasing the number of CVs. The high increase in execution time means the MOO algorithm is impractical to be applied in the gateway selection system, which needs to be executed in real-time, especially in an urban area where the vehicles' density is high.

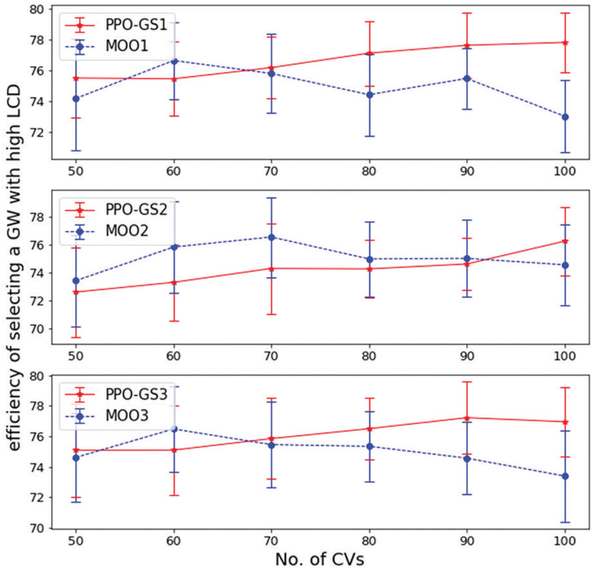

Fig. 8. The percentage of the gateway selection with the highest LCD.

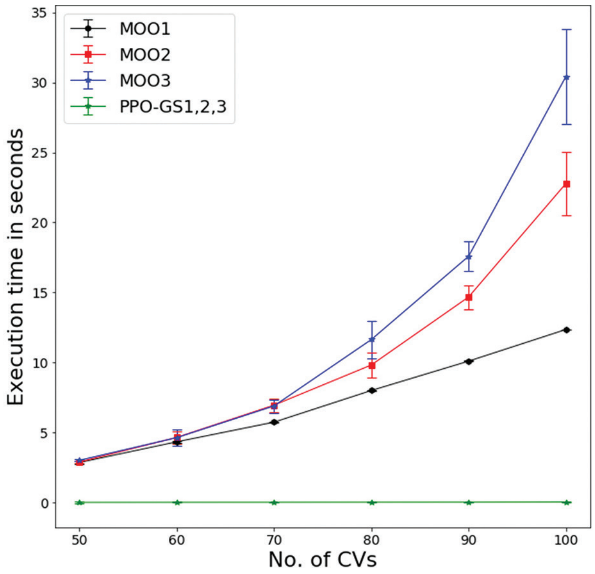

Fig. 9. Execution time.

5. CONCLUSION AND FUTURE WORK

In this article, a new gateway selection algorithm is presented with the aim of finding the best mobile gateway for vehicles in need of Internet access. For this purpose, an integrated system is proposed using two cloud servers. The first one collects all necessary information about CVs and MGs, while the second one, where the gateway discovery system resides, uses the data collected by the first server to train the agent. RL uses two objectives to optimize the gateway discovery system. These two objectives are the link connectivity duration between the vehicles in need of the Internet and the gateways and bandwidth limitation of the gateways. The weighted sum approach is employed to find a trade-off between the two contradicting objectives. The proposed algorithm uses the proximal policy optimization strategy to implement and train the agent. Different agents are created based on the objectives' preferences. Compared with the existing mobile gateway selection algorithms, the simulation results show that the proposed approach is effective in terms of increasing the number of connected vehicles, distributing the traffic among gateways, and reducing the execution time. In future work, we plan to add more parameters to the agent state to make it more expressive, like road density and the number of neighbors for each gateway.

6. REFERENCES

[1] M. Lee, T. Atkison, "VANET applications : Past, present, and future", Vehicular Communications, Vol. 28, 2021, p. 100310.

[2] H. Alabbas, Á. Huszák, "A New Clustering Algorithm for Live Road Surveillance on Highways based on DBSCAN and Fuzzy Logic", International Journal of Advanced Computer Science and Applications, Vol. 11, No. 8, 2020, pp. 580–587.

[3] F. Cunha et al., "Data communication in VANETs: Protocols, applications, and challenges" Ad Hoc Networks, Vol. 44, pp. 90–103, 2016.

[4] Y. Fangchun, W. Shangguang, L. I. Jinglin, L. I. U. Zhihan, S. U. N. Qibo, "An Overview of Internet of Vehicles", China Communications, Vol. 11, no. 10, pp. 1-15, 2014.

[5] P. Anusha, S. K. Shabanabegum, R. Pavaiyarkarasi, E. Seethalakshmi, and K. Vadivukkarasi, "Smart internet of vehicle maintenance system", Materials Today: Proceedings, 2020.

[6] M. Alawi, E. Sundararajan, "Gateway Selection Techniques in Heterogeneous Vehicular Network: Review and Challenges ", Proceedings of the IEEE 6th International Conference on Electrical Engineering and Informatics, Langkawi, Malaysia, 25-27 Nov. 2017, pp. 1-6.

[7] S. Sharma, B. Kaushik, "A survey on internet of vehicles : Applications, security issues & solutions", Vehicular Communications, Vol. 27, 2021, p. 100289

[8] Y. Lin, J. Shen, H. Weng, "Gateway Discovery in VANET Cloud", Proceeding of the IEEE International Conference on High Performance Computing and Communications, Banff, AB, Canada, 2-4 September 2011, pp. 951-954.

[9] M. H. Badole, T. Raju, "Protocol design for an efficient Gateway Discovery & Dispatching for Vehicular Ad Hoc Network", International Journal of Application or Innovation in Engineering & Management, Vol. 3, No. 2, 2014, pp. 117–120.

[10] R. Ghebleh, "A comparative classification of information dissemination approaches in vehicular ad hoc networks from distinctive viewpoints: A survey", Computer Networks, Vol. 131, 2018, pp. 15-37,

[11] R. Hussain, J. Son, H. Eun, S. Kim, H. Oh, "Rethinking Vehicular Communications: Merging VANET with Cloud Computing", Proceeding of the 4th IEEE International Conference on Cloud Computing Technology and Science, Taipei, Taiwan, 3-6 December 2012, pp. 606-609.

[12] M. Gerla, E. Lee, G. Pau, U. Lee, "Internet of vehicles: From intelligent grid to autonomous cars and vehicular clouds", Proceeding of the IEEE World Forum on Internet of Things, Seoul, Korea (South), 6-8 March 2014, pp. 241-246.

[13] M. Chaqfeh, N. Mohamed, I. Jawhar, Jie Wu, "Vehicular Cloud data collection for Intelligent Transportation Systems", Proceeding of the IEEE 3rd Smart Cloud Networks & Systems, Dubai, United Arab Emirates, 19-21 December 2016, pp. 1-6.

[14] S. Bitam, A. Mellouk, S. Zeadally, "VANET-cloud: a generic cloud computing model for vehicular Ad Hoc networks", IEEE Wireless Communications, Vol. 22, No. 1, 2015, pp. 96-102.

[15] D. Abada, A. Massaq, A. Boulouz, M. Salah, "An Adaptive Vehicular Relay and Gateway Selection Scheme for Connecting VANETs to Internet via 4G LTE Cellular Network", Proceeding of the IEEE International Conference of Computer Science and Renewable Energies, Agadir, Morocco, 22-24 July 2019, pp. 1-8.

[16] G. Zhioua, H. Labiod, N. Tabbane, S. Tabbane, "FQGwS: A gateway selection algorithm in a hybrid clustered VANET LTE-advanced network: Complexity and performances", Proceeding of the IEEE International Conference on Computing, Networking and Communications, Honolulu, HI, USA, 3-6 February 2014, pp. 413-417.

[17] B. Sharef, R. Alsaqour, M. Alawi, M. Abdelhaq, "Robust and trust dynamic mobile gateway selection in heterogeneous VANET-UMTS network", Vehicular Communications, Vol. 12, 2018, pp. 75–87.

[18] A. Idrissi, S. Retal, H. Rehioui, A. Laghrissi, "Gateway selection in Vehicular Ad-hoc Network", Proceeding of the IEEE 5th International Conference on Information & Communication Technology and Accessibility, Marrakech, Morocco, 21-23 December 2015, pp. 1-5.

[19] S. Retal, A. Idrissi, "A multi-objective optimization system for mobile gateways selection in vehicular Ad-Hoc networks", Computers and Electrical Engineering, Vol. 73, 2019, pp. 289–303.

[20] G. Setiwan, S. Iskander, S. S. Kanhere, Q. J. Chen, "Feasibility Study of Using Mobile Gateways for Providing Internet Connectivity in Public Transportation Vehicles", Proceedings of the International Conference on Wireless Communications and Mobile Computing, British Columbia, Canada, July 3-6, 2006, pp. 1097–1102.

[21] R. S. Sutton, A. G. Barto, "Reinforcement learning: An introduction", 2nd Edition, MIT Press, 2018.

[22] H. Dong, H. Dong, Z. Ding, S. Zhang, Chang, "Deep Reinforcement Learning", Springer, 2020.

[23] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, "Proximal policy optimization algorithms", arXiv Prepr. arXiv1707.06347, 2017.

[24] C. Liu, X. Xu, D. Hu, "Multiobjective reinforcement learning: A comprehensive overview", IEEE Transactions on Systems, Man, and Cybernetics: Systems, Vol. 45, No. 3, 2014, pp. 385–398.

[25] A. Raffin, A. Hill, M. Ernestus, Gleave, A. Gleave, A, Kanervisto, N. Dormann, Stable Baselines3, https://github.com/DLR-RM/stable-baselines3 (accessed: 2021).