# Faster Convergence of Q-Learning in Cognitive Radio-VANET Scenario

**Mohammad Asif Hossain, Rafidah Md Noor, Saaidal Razalli Azzuhri, Muhammad Reza Z'aba, Ismail Ahmedy, Shaik Shabana Anjum, Wahidah Md Shah and Kok-Lim Alvin Yau**

**Abstract** Cognitive Radio (CR) based Vehicular Ad hoc Network (VANET) or CR-VANET has become a very promising research domain. VANET is used to reduce road accidents, traffic congestion, and to provide other user experiences such as uninterrupted entertainment services. CR, on the other hand, solves bandwidth scarcity issue of VANET. For the high-speed mobility of the vehicles, the cognitive process of CR faces several challenges. Machine Learning (ML) has arrived as an integral tool to handle such challenges. Q-learning algorithm, a member of Reinforcement Learning (RL), which is a type of ML, is the most suitable for CR-VANET as it does not need any prior environment model and training dataset. But the problem is that it takes a longer time for learning purposes. In this paper, a dynamic ML framework is proposed. Case-based reasoning learning, cooperative spectrum sensing, teacher-student transfer learning approach will be aligned with the Q-learning for the faster convergence regarding the spectrum sensing issues in CR-VANET. The framework will accelerate the learning of the vehicles, and that is very important for the energy-efficient and real-life VANET implementation.

**Keywords** VANET · Cognitive radio · Reinforcement learning · Transfer learning · Q-learning · Case-based reasoning

M. A. Hossain · R. M. Noor (✉) · S. R. Azzuhri · M. R. Z'aba · I. Ahmedy · S. S. Anjum
Faculty of Computer Science and Information Technology, University of Malaya,
Kuala Lumpur, Malaysia
e-mail: fidah@um.edu.my

R. M. Noor
Centre for Mobile Cloud Computing Research, University of Malaya,
Kuala Lumpur, Malaysia

W. M. Shah
Faculty of Information and Communication Technology, Universiti Teknikal Malaysia
Melaka, Melaka, Malaysia

K.-L.A. Yau
School of Science and Technology, Sunway University, Selangor, Malaysia

Z. Zakaria and R. Ahmad (eds.), *Advances in Electronics Engineering*,
Lecture Notes in Electrical Engineering 619,
https://doi.org/10.1007/978-981-15-1289-6_16

# 1 Introduction

Every year, around 1.25 million people die from the road accidents [1] and the resulting congestion incurs a huge amount of money (In the U.S. alone, congestion cost $305 billion in 2017 [2]). VANET has emerged as a solution to these alarming situations. For implementing VANET, a gigantic amount of real-time data (such as of GPS (Global Positioning System), radar, camera, LIDAR (Light Detection and Ranging), Sonar, sensors data), and infotainment data will be exchanged in the coming years. According to Intel, each smart vehicle is going to generate and consume approximately 4 terabytes of data in on average per day driving by 2020 [3].

IEEE 802.11p or IEEE 1609, also known as Dedicated Short-Range Communications (DSRC) standard is reserved for the vehicular networks with 75 MHz bandwidth in the frequency range of 5.85 to 5.925 GHz. This allocated bandwidth is not sufficient enough to accommodate such a massive amount of data [4]. On the other hand, licensed bandwidths or frequencies such as TV band or military radio band are not properly utilized [5]. The report shows that more than 60% bandwidth of below 6 GHz spectrum is not being used or not properly utilized [6]. CR, the concept coined by Mitola & Maguire [7], has emerged as the solution in the bandwidth scarcity problem. CR users are allowed to sense and use these underutilized licensed channels dynamically in an opportunistic manner, as well as for spectrum mobility that allows users to vacate licensed channels re-occupied by licensed users (primary users or PUs). The latency for the safety message exchange must be lower than 100 ms, but in general, the cognitive processes takes around 2 s time [8]. Moreover, a huge amount of network overhead is transmitted due to such cognitive processes and unnecessary repetitive computational tasks have to be performed. These will lead to unnecessary energy consumption.

ML can be applied in CR-VANET to make it more intelligent to adapt the uncertain radio environment to solve those issues. It can ensure faster decision, reliability, energy efficiency, and enhanced QoS [9]. There are three main categories of ML techniques: supervised learning, unsupervised learning, and RL. Other learning methods such as semi-supervised learning, online learning, and transfer learning are the variation of these three categories [10]. Q-learning, a type of among several RL algorithms, is found as the most suitable for the CR-VANET scenario due to its adaptability with the dynamic environment, model-free requirement, and working capability without training dataset [11]. Here, agents face an unpredictable environment by selecting appropriate actions by using mathematical approaches and receiving rewards consequently. The main issue faced by a Q-learning agent is that it takes longer learning phases, i.e. a huge number of iterations are required for the convergence. This is due to its learning itself all alone. In this paper, a dynamic learning framework has been proposed. The objective of this framework is to reduce the overall learning time of the vehicles about the vacant spectrums on the surrounded environment. The idea of teacher-student approach (a type of transfer learning which is a feature of ML) along with case-based reasoning (CBR) will accelerate the learning time of Q-learning. In a teacher-student approach, an already learned vehicle (teacher) will share its own sensing information to the learning vehicle (student) [12]. CBR, another type of ML, tries to solve new problems by reusing past solutions that were used to solve similar problems. This cognitive process uses prior stored 'case' (results and experience) to fit a new similar problem situation [8].

The remainder of this paper is organized as follows: Sect. 2 discusses the related works, Sect. 3 provides the overview and the problem formulation of Q-learning, Sect. 4 discusses the proposed framework, Sect. 5 describes the performance evaluation methods and parameters, and finally, Sect. 6 concludes the paper.

## 2 Related Works

Several works were done in the fields of spectrum sensing in CR-VANETs by using Q-learning. In [13], the author proposed architecture by using Q-learning and CBR for VANET to enable automatic learning of the radio environment by the vehicles. The authors in [14] showed that by using this learning, the total energy consumption due to the spectrum sensing can be reduced to only about 1.72% compared to the traditional spectrum sensing method. In [15], the authors used deep Q-learning for designing an optimal data transmission scheduling scheme in CR-VANET to minimize transmission costs. They used cache memory for taking the decision. Their scheme's convergence took place after 13,000 to 20,000 iterations at 28 m/s vehicle speed. Morozs et al. in [16] proposed a scheme, which integrates distributed Q-learning and CBR aimed to facilitate a number of learning processes running in parallel. They got the best result after 1,000,000 iterations. RL method was considered for the CR network with RF energy harvesting in [17]. Their proposed scheme was for the optimum switching between the transmit mode, energy harvesting mode, and receiving mode of the CR users. They got average throughput converges to 0.68 after 1,000,000 iterations.

The above-mentioned works found very good performance in spectrum sensing in terms of higher probability of PUs detection with a lower probability of false alarm, but with a very slow convergence rate. They needed a huge number of iteration to learn the environment optimally. For the practical point of view, these learning times are quite infeasible. Authors in [18] gave some insights about the way to accelerate Q-learning time, though it was theoretical and was not considered the aspects of CR-VANET. This paper is targeted to reduce such learning time (i.e. make the convergence faster).

## 3 Q-Learning Algorithm

Q-learning, the most used type of RL, is an on-line algorithm, which enables an agent to learn in an interactive manner with its surrounding environment. The main aim of Q-learning is to exploit the long-term rewards receiving in the future. It does not require any environment model and dataset for the training. In Q-learning, an agent or the learner (say a CR based vehicle) is interacting with the radio environment (comprising everything outside the agent).

From Fig. 1, it is shown that, at each step $t$, the agent observes the state of its surrounding environment $s_t \epsilon S$, where $S$ is a set of possible states. Based on knowledge gained at $s_t$, the agent selects an action $a_t \epsilon A$, where $A$ is a set of actions. At the next step $t + 1$, the environment transits to a new state $s_{t+1}$ and the agent gets a reward of $r_t$. Based on the reward table, the agent chooses the next action (it may be beneficial or may be harmful) and then they update a new value called Q-value mapping of state-action pairs Q ($s_t$, $a_t$). Several Q-values are stored in the Q-table. For example, in CR-VANET scenario, an action might be choosing any spectrum for accessing, the state might be the location and time of the vehicle. If the sensed spectrum faces interferences by the PUs, the agent would get a negative reward, otherwise gets a positive reward.

After every action, the agent gets the reward and updates its Q-value based on Eq. (1).

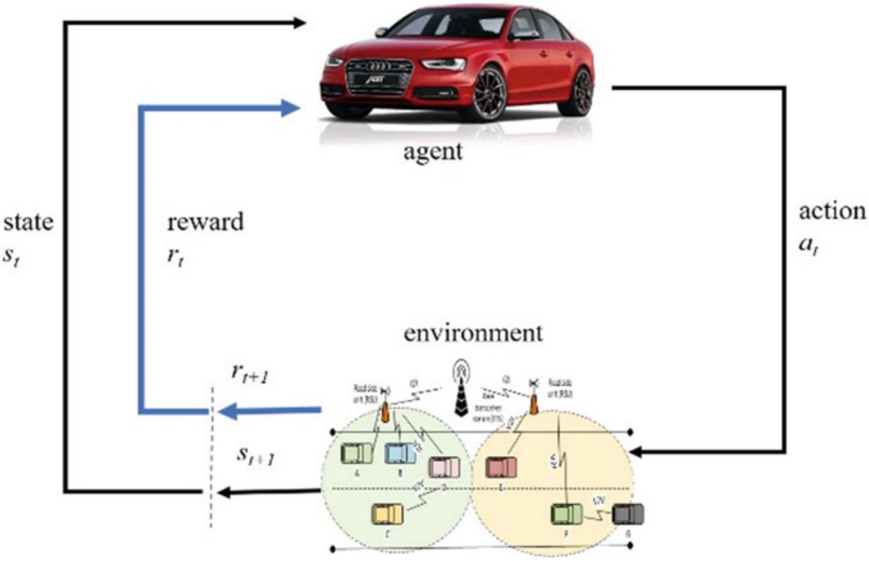


**Fig. 1** Q-learning approach

$$\begin{aligned} Q_{new}(state, action) \leftarrow & (1-\alpha)Q_{old}(state, action) \\ & + \alpha(reward + \gamma \, max \, Q_{old}(next\, state, all\, actions)) \end{aligned}$$

$$\because \quad Q_{t+1}(s_t, a_t) \leftarrow (1-\alpha)Q_t(s_t, a_t) + \alpha\left[r_{t+1}(s_{t+1}, a_t) + \gamma \max_{a \in A} Q_t(s_{t+1}, a)\right] \quad (1)$$

Here,

$\alpha$: The *learning rate*, which determines how much the new Q-value overrides the previous Q-value. $\alpha$ ranges from 0 to 1. The higher value of $\alpha$ means the higher speed of the learning process (may lead to faster convergence), but sometimes stability is lost and failed to converge. The lower the value of $\alpha$, smoother the learning process but slower rate of convergence.
$\gamma$: The *discount factor*, which implies how much importance is given to future rewards.
$r$: The *reward* received by the agent. The short-term reward is called the *delayed reward* and the future reward is called the *discounted reward.*

There are two policies for taking action. When the agent chooses for exploitation (uses existing knowledge to select the best action), it uses an optimal policy and when it chooses for exploration (needs more knowledge), it uses a random policy. The agent receives positive delayed rewards when it selects a proper action for a particular state. Positive value increases and the respective Q-value, and vice versa [19]. Therefore, the target of Q-learning is to get an optimal policy (agent behavior) $\pi$*:* $S \rightarrow A$, which can maximize the reward at state $S$ [20].

The optimal Q-value for a particular state can be written as:

$$V^{\pi^*}(s_t) = \max_{a \in A} Q_t(s_t, a) \quad (2)$$

Therefore, the optimal policy can be written as:

$$\pi^*(s_t) = \arg\max_{a \in A} Q_t(s_t, a) \quad (3)$$

From the above discussions, it is clear that the convergence rate depends on the quality of Q-table and the value of $\alpha$ and $\gamma$. The more reward an agent accumulated, the better Q-table would get, and therefore, the convergence will be faster. But the issue is Q-learning algorithm is learning totally by itself, not taking any helps from others. For better performance and convergence, it has to face the tradeoff between exploration and exploitation. More exploration provides better decision (sacrifices immediate rewards hoping for more future rewards), but slower convergence, on the other hand, quick exploitation might provide faster convergence, but poorer performance. If the Q-table is updated with more rewarded state-action pairs, overall convergence would be faster.

## 4 Proposed Dynamic Machine Learning Framework

In this paper, a dynamic ML framework that includes Q-learning and CBR has been developed. Teacher-student transfer learning approach has also been used in the framework. Here, a learned vehicle (teacher) shares its own sensing information to the learning vehicle (student) [12]. The vehicle chooses the best ML based on the proposed framework. Suppose, if the user chooses the same known path at the same time of the day, the CBR would be used, and if the environment is unknown to the CBR-database, Q-learning method would be used.

In this proposed theme, like teacher-student approach in [12], a learned vehicle, for example, might have the best action-state pair or best Q($s_t$, $a_t$). If the learning vehicle is getting Q value from this learned vehicle, it does not need additional exploration for that state. For example, in Fig. 2, a learning car (A) has broadcasted a request for the spectrum sensing information to the neighbor vehicles. A is in say $s_{tk}$ state, on the request it will include this state value. A teacher (say B) has the best-rewarded information regarding state $s_{tk}$, so it will then forward Q($s_{tk}$, $a_{tk}$) as the response to A. Another car (C) say, for example, does not have information regarding the $s_{tk}$ state. So, it would not respond. After getting the Q($s_{tk}$, $a_{tk}$) from B, the car A will keep this Q value to its Q-table. So, in future, when car A is in the same state ($s_{tk}$), it would not go for exploration state. In this way, by cooperation, a learning vehicle can increase its learning process. Figure 3 shows the proposed framework. This framework will provide faster convergence and reduce the overall sensing time, hence, provides the energy efficient and improved QoS CR-VANET. This framework is also described in Algorithm 1 and the Q-learning algorithm in Algorithm 2. Here, when a vehicle selects the destination and starts its journey, it will search its own database whether the route (the road) is already known or unknown. The $i_{th}$ vacant spectrum information contains the location ($l_i$), time ($t_i$), and channel ($c_i$). If the information is found known by the searching database, it retrieves spectrum information from the database (learned previously) and uses that vacant channel.

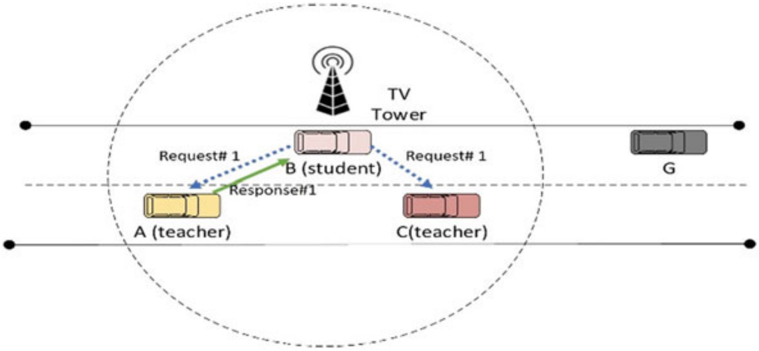


**Fig. 2** Teacher-student transfer learning approach

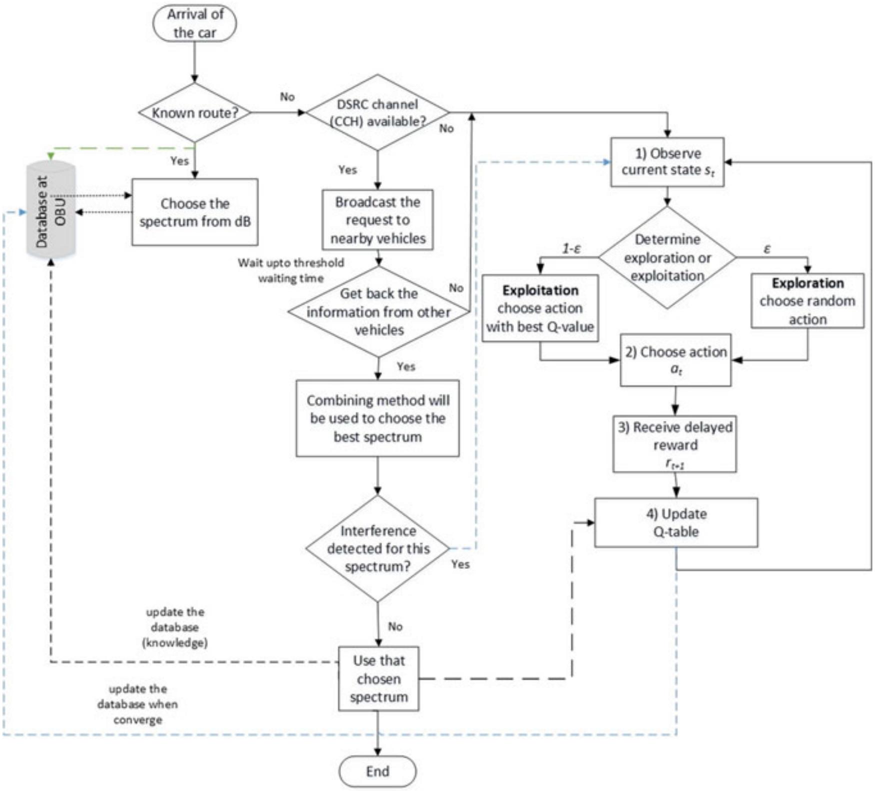
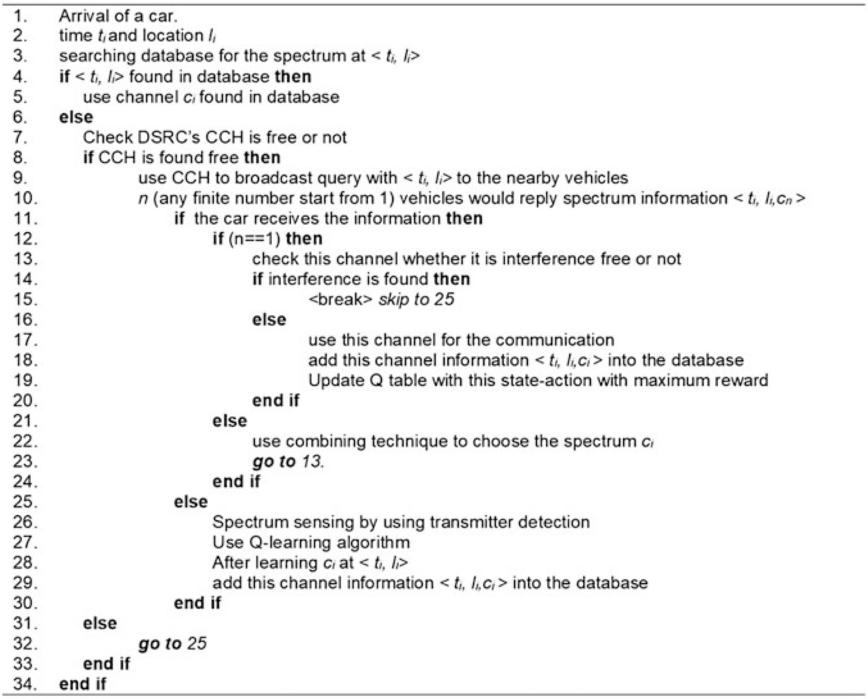

Arrival of the car
Known route?
No
DSRC channel (CCH) available?
No
Yes
Database at OBU
Choose the spectrum from dB
Yes
Broadcast the request to nearby vehicles
Wait upto threshold waiting time
Get back the information from other vehicles
No
Yes
Combining method will be used to choose the best spectrum
Interference detected for this spectrum?
Yes
No
update the database (knowledge)
Use that chosen spectrum
update the database when converge
End
1) Observe current state $s_t$
Determine exploration or exploitation
1-ε
ε
Exploitation choose action with best Q-value
Exploration choose random action
2) Choose action $a_t$
3) Receive delayed reward $r_{t+1}$
4) Update Q-table

**Fig. 3** Proposed framework of dynamic machine learning in CR-VANET

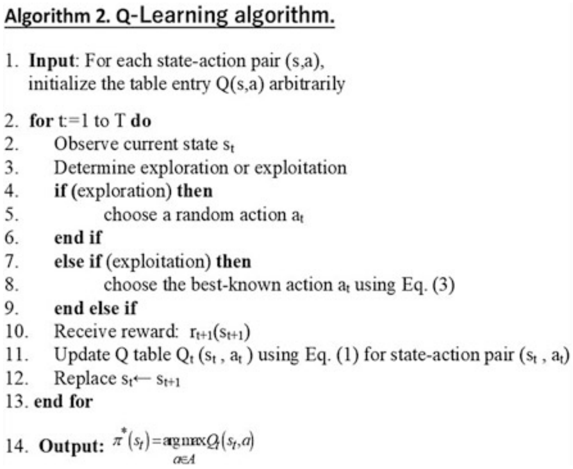
If the route is found unknown, the vehicle will look for whether the DSRC's common control channel (CCH) is available or not. If it's found available, it broadcasts a query message to its neighbor vehicles for learning the free channel of that route on that time. If the vehicle gets the responses from several vehicles, it will use any suitable combining method (like maximal-ratio combining or MRC) to choose the best channel. It will then test whether the channel is really free or not by using any detection method. If it finds interference-free, it will use that channel and stores this information ($l_i$, $t_i$, $c_i$) to the database and to the Q-table. If the vehicle finds CCH unavailable or detects interference or does not get information from any vehicle, it will go for non-cooperative spectrum sensing by using a primary transmitter detection method. Q-learning will be used for taking further action and get backs as rewards/punishments. Here, the action means selecting the spectrum to use, the agent gets a reward when it finds interference-free (absence of PUs) and gets punishment when it finds interference on its chosen spectrum. After some iterations, it will be converged and then updates the database. It will add ($l_i$, $t_i$, $c_i$) into the database. For the Q-learning, the $\varepsilon$-greedy policy has been considered.

In, ε-greedy policy, the agent chooses exploration with a small probability ε and exploitation with probability (1 − ε).

**Algorithm 1 Dynamic Spectrum Sensing**



```
1.  Arrival of a car.
2.  time $t_i$ and location $l_i$
3.  searching database for the spectrum at < $t_i$, $l_i$>
4.  if < $t_i$, $l_i$> found in database then
5.      use channel $c_i$ found in database
6.  else
7.      Check DSRC's CCH is free or not
8.      if CCH is found free then
9.              use CCH to broadcast query with < $t_i$, $l_i$> to the nearby vehicles
10.             $n$ (any finite number start from 1) vehicles would reply spectrum information < $t_i$, $l_i$,$c_n$ >
11.                 if the car receives the information then
12.                     if (n==1) then
13.                         check this channel whether it is interference free or not
14.                         if interference is found then
15.                                 <break> skip to 25
16.                         else
17.                                 use this channel for the communication
18.                                 add this channel information < $t_i$, $l_i$,$c_i$ > into the database
19.                                 Update Q table with this state-action with maximum reward
20.                         end if
21.                     else
22.                         use combining technique to choose the spectrum $c_i$
23.                         go to 13.
24.                     end if
25.                 else
26.                     Spectrum sensing by using transmitter detection
27.                     Use Q-learning algorithm
28.                     After learning $c_i$ at < $t_i$, $l_i$>
29.                     add this channel information < $t_i$, $l_i$,$c_i$ > into the database
30.                 end if
31.     else
32.             go to 25
33.     end if
34. end if
```



**Algorithm 2. Q-Learning algorithm.**

```
Input: For each state-action pair (s,a),
initialize the table entry Q(s,a) arbitrarily
for t:=1 to T do
    Observe current state $s_t$
    Determine exploration or exploitation
    if (exploration) then
        choose a random action $a_t$
    end if
    else if (exploitation) then
        choose the best-known action $a_t$ using Eq. (3)
    end else if
    Receive reward: $r_{t+1}(s_{t+1})$
    Update Q table $Q_t$ ($s_t$ , $a_t$ ) using Eq. (1) for state-action pair ($s_t$ , $a_t$)
    Replace $s_t \leftarrow s_{t+1}$
end for
Output: $\pi^*(s_t) = \arg\max_{a \in A} Q_t(s_t, a)$
```

The overall learning will not be solely depending on previous data, not on CSS only nor on Q-learning only. The proposed learning is a kind of hybrid learning and will eventually become faster and more reliable. The main concept here is that, at first step, every vehicle searches its own database, if it finds, it uses that vacant frequency. If it does not find information from its own database, it would ask for help from other surrounding vehicles. The vehicles that already know about the sensing information (on that route and on that time) will deliver their learned sensing information to that vehicle. Here, the teacher-student transfer learning approach has been used. After getting the information from several vehicles, the vehicle would use fusion or combining technique. Learned sensing information will be fed to the database and to the Q-table. If all these stages fail, it will go for the non-cooperative SS and RL phases. This system provides faster spectrum information and increases the convergence rate.

## 5 Performance Evaluation

For the performance evaluation purposes, SUMO (Simulation of Urban MObility) simulator, Network simulator 3 (NS3), and Python programming will be used. Figure 4 shows the steps of getting the results for the performance evaluation. By using SUMO, the real-life mobility model and VANET will be designed, then this will be integrated with the NS3 to add the feature of CR. After running the simulation, spectrum data of CR-VANET would be obtained. These data will be fed to the Python, wherein the framework of Q-learning, CBR, teacher-student, and CSS would be implemented. After performing data analysis by Python, the results would be obtained.

Following performance metrics would be used for the performance measurements of the proposed framework:

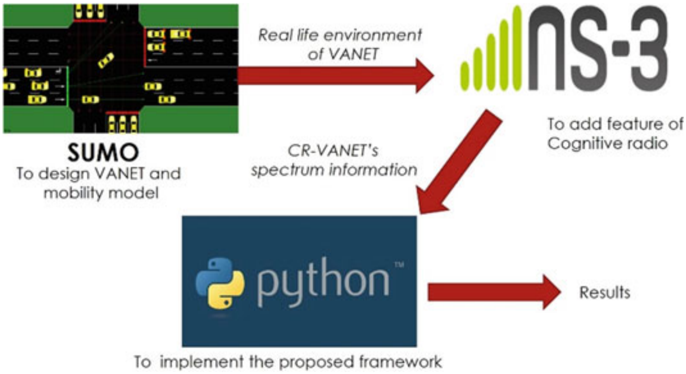

SUMO
To design VANET and mobility model
Real life environment of VANET
NS-3
To add feature of Cognitive radio
CR-VANET's spectrum information
python
Results
To implement the proposed framework

**Fig. 4** Steps in performance evaluation

*Convergence rate*: it defines how fast an agent (vehicle) learns the surrounded complex environment or simply the number of iterations needed for an algorithm to start providing the best optimal value. The faster the convergence rate (less iteration required), the better the algorithm performs. The aim of this paper is to make the convergence faster (learns the system within a short period of time).

*The probability of false alarm versus the probability of detection*: the probability of false alarm means the probability of declaring about the presence of a PU, though that sensed spectrum is not really occupied by any PU. On the other hand, the probability of detection represents the probability declaring the presence of a PU and that sensed spectrum is truly occupied by that PU. In CR-VANET, it is one of the most used performance metrics.

*Energy efficiency*: It is the measurement by which the performance of a system can be evaluated. When a system provides the same services but with less energy consumption compared to other systems, then it can be said that the earlier system performs better in terms of energy and it is an energy efficient system.

*Delay*: It is one of the major issues in CR-VANET scenario. It is the difference between the theoretical time taken by a system and the actual time it takes to perform any task. The cognitive process should be performed with a very lower delay. High delay reduces the overall performance of a system.

This research work is expecting a higher probability of detection and lower probability of false alarm, faster convergence, higher energy efficiency, and lesser delay compared to the existing techniques and methods for the spectrum sensing in CR-VANET scenario.

## 6 Conclusion

Vehicular Ad hoc Network (VANET) has emerged as one of the major solutions to enhance road safety, reduce traffic congestion, and improve quality-of-service (QoS). Cognitive Radio (CR), on the hand, has appeared to alleviate the spectrum scarcity issue of exponentially growing VANETs. Machine learning tools are now becoming an integral part of CR-VANET to boost its advantages. In this paper, a dynamic machine learning framework has been proposed. The framework consists Q-learning, case-based reasoning, and teacher-student transfer learning concept. The proposed framework is expected the improvement in terms of convergence rate. The proposed method is also expected to provide reliable learning to the vehicles in very dynamic environments with reduced delay and network overhead. In future work, we will analyze, design, and validate this proposed dynamic machine learning framework with considering various challenges such as PU activity models, hidden PUs problem, Doppler effects and so on.

**Acknowledgements** This work was supported by the Partnership Grant under Grant RK004-2017 and Grant CR-UM-SST-DCIS-2018-01 between University of Malaya and Sunway University.

## References

1. WHO (2015) Road safety report 2015 (Online). Available: http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/
2. Schneider B (2018) Traffic's mind-boggling economic toll (Online). Available: https://www.citylab.com/transportation/2018/02/traffics-mind-boggling-economic-toll/552488/
3. Winter K (2017) For self-driving cars, there's big meaning behind one big number: 4 terabytes (Online). Available: https://newsroom.intel.com/editorials/self-driving-cars-big-meaning-behind-one-number-4-terabytes/
4. Singh KD, Rawat P, Bonnin J-M (2014) Cognitive radio for vehicular ad hoc networks (CR-VANETs): approaches and challenges. EURASIP J Wirel Commun Netw 2014(1):49
5. Vo QD, Choi JP, Chang HM, Lee WC (2010) Green perspective cognitive radio-based M2M communications for smart meters. In: 2010 international conference on information and communication technology convergence, ICTC 2010, pp 382–383
6. Spectrum Policy Task Force Report: FCC, in Cambridge University Press, 2012, 2002
7. Mitola G, Maguire J (1999) Cognitive radio: making software radios more personal. IEEE Pers Commun 6(4):13–18
8. Chen S, Vuyyurut R, Altintas O, Wyglinski AM (2011) Learning in vehicular dynamic spectrum access networks : opportunities and challenges. In: 2011 International Symposium on Intelligent Signal Processing and Communications Systems, pp 5–10
9. Yau KLA, Komisarczuk P, Teal PD (2010) Applications of reinforcement learning to cognitive radio networks. In: 2010 IEEE International Conference on Communications Workshop, pp 1–6
10. Liang L, Ye H, Li GY (2018) Towards intelligent vehicular networks: a machine learning framework. IEEE Internet Things J 6(1)
11. Wu C, Ohzahata S, Kato T (2013) Flexible, portable, and practicable solution for routing in VANETs: a fuzzy constraint Q-Learning approach. IEEE Trans Veh Technol 62(9):4251–4263
12. Da Silva FL, Glatt R, Costa AHR (2017) Simultaneously learning and advising in multiagent reinforcement learning. In: Proceedings of the 16th conference on autonomous agents and multiagent systems, pp 1100–1108
13. Chen S, Vuyyuru R, Altintas O, Wyglinski AM (2011) On optimizing vehicular dynamic spectrum access networks: automation and learning in mobile wireless environments. In: IEEE vehicular networking conference, VNC, pp 39–46
14. Grace D, Chen J, Jiang T, Mitchell PD (2009) Using cognitive radio to deliver 'green' communications. In: Proceedings 2009 4th international conference on cognitive radio oriented wireless networks and communications, CROWNCOM 2009, pp 2–7
15. Zhang K, Leng S, Peng X, Pan L, Maharjan S, Zhang Y (2018) Artificial intelligence inspired transmission scheduling in cognitive vehicular communications and networks. IEEE Internet Things J 6(2)
16. Morozs N, Clarke T, Grace D (2016) Cognitive spectrum management in dynamic cellular environments: a case-based Q-learning approach. Eng Appl Artif Intell 55:239–249
17. Van Huynh N, Hoang DT, Nguyen DN, Dutkiewicz E, Niyato D, Wang P (2018) Reinforcement learning approach for RF-powered cognitive radio network with ambient backscatter. In: CoRR, vol abs/1808.0, pp 1–6
18. Potapov A, Ali MK (2003) Convergence of reinforcement learning algorithms and acceleration of learning. Phys Rev E 67(2):26706
19. Ling MH, Yau K-LA, Qadir J, Poh GS, Ni Q (2015) Application of reinforcement learning for security enhancement in cognitive radio networks. Appl Soft Comput J 37:809–829
20. Sutton RS, Barto AG (2017) Reinforcement learning: an introduction, 2nd edn. The MIT Press