Exploring Cooperative Multi-agent Reinforcement Learning Algorithm (CMRLA) for Intelligent Traffic Signal Control

Deepak A. Vidhate[1(✉)] and Parag Kulkarni[2]

[1] Department of Computer Engineering, College of Engineering, Pune, Pune, Maharashtra, India
dvidhate@yahoo.com
[2] iKnowlation Research Lab. Pvt. Ltd., Pune, Maharashtra, India
parag.india@gmail.com

Abstract. Traffic crisis frequently happen because of traffic burden by the large number automobiles are on the path. Increasing transportation move and decreasing the average waiting time of each vehicle are the objectives of cooperative intelligent traffic control system. Each signal wishes to catch better travel move. During the course, signals form a strategy of cooperation in addition to restriction for neighboring signals to exploit their individual benefit. A superior traffic signal scheduling strategy is useful to resolve the difficulty. The several parameters may influence the traffic control model. So it is hard to learn the best possible result. Traffic light controllers are not expert to study from previous results. Due to this they are unable to include uncertain transformation of traffic flow. Reinforcement learning algorithm based traffic control model can be used to obtain fine timing rules by properly defining real time parameters of the real traffic scenario. The projected real-time traffic control optimization prototype is able to continue with the traffic signal scheduling rules successfully. The model expands traffic value of the vehicle, which consists of delay time, the number of vehicles stopped at signal, and the newly arriving vehicles to learn and establish the optimal actions. The experimentation outcome illustrates a major enhancement in traffic control, demonstrating the projected model is competent of making possible real-time dynamic traffic control.

Keywords: Cooperation schemes · Intelligent traffic control
Reinforcement learning

1 Introduction

Large number of vehicles dispersed in a large and board urban area. This makes a difficult and complicated work to successfully take care of such a large scale, dynamic, and distributed system with a high degree of uncertainty [1]. Though the number of vehicles are getting more and more in major cities, most of the current traffic control methods have not taken benefit of a intelligent control of traffic light [2]. It is observed that sensible traffic control and enhancing the deployment effectiveness of roads is an efficient and cost effective technique to resolve the urban traffic crisis in majority urban

A. V. Deshpande et al. (Eds.): SmartCom 2017, CCIS 876, pp. 71–81, 2018.
https://doi.org/10.1007/978-981-13-1423-0_9

areas [3]. Major vital part of intelligent transportation system is traffic signal lights control strategy becomes necessary [4]. There are so various parameters that have an effect on the traffic lights control. Static control method is not feasible for rapid and irregular traffic flow. The paper suggests a dynamic traffic control framework which is based on reinforcement learning [5]. The reinforcement learning can present a very crucial move to resolve the above cited problems. It is effectively deployed in resolving various problems [6]. The framework defines different traffic signal control types as action selections; the number of vehicles arriving and density of vehicle at a junction are observed as environment condition. Signal management parameters, like delay time, the number of stopped vehicles, and the total vehicle density are described as received rewards.

The article is described in four parts. Section 2 describes about the traffic estimation parameters. Cooperative multi-agent reinforcement learning algorithm (CMRLA) is proposed in Sect. 3. Section 4 discuss about the system model, including definitions pertaining the state, action, and reward function. Section 5 discuss about experiment and analysis of the results followed by concluding remark.

2 Traffic Estimation Parameters

In traffic management a very crucial responsibility is handled by signal lights control. A practical time allotment method ensures that in usual conditions the traffic moves seamlessly. Normally applied traffic estimation parameters [7] comprises of delay time, the number of automobiles stopped at intersection, and number of newly arriving automobiles.

2.1 Delay Time

The delay between the real time and theoretically calculated time for a vehicle to leave a signal is defined as delay time. In practice, we can get total delay time during a certain period of time and average delay time of a cross to evaluate the time difference. The more delay time indicates the slower average speed of a vehicle to leave a signal.

2.2 Number of Vehicles Stopped

How many vehicles are waiting behind stop line to leave the road signal gives the number of vehicles stopped. The indicator [8] is used to measure the smooth degree of road as well as the road traffic flow. It is defined as

$$\text{stop} = \text{stopG} + \text{stopR} \tag{1}$$

where stopR is the number of automobiles stopped before the red light and stopG is the number of automobiles stopped before the green light.

2.3 Number of Vehicles Newly Arrived

The ratio of the actual traffic flow to the maximum available traffic flow gives the signal saturation. Newly arrived vehicle is calculated as

$$S = \frac{traffic\ flow}{(dr * sf)} \tag{2}$$

where s_f is traffic flow of the signal and d_r is the ratio of red light duration to green light duration.

2.4 Traffic Flow Capacity

Highest number of vehicles crossing through the signal is shown by traffic flow capacity. The result of signal control strategy is given by the indicator. Traffic signal duration and traffic flow capacity are associated with each other. Generally more signal crossing capability is a result of more crossing period.

3 Cooperative Multi-agent Reinforcement Learning Algorithm (CMRMA)

Synchronization in multi-agent reinforcement generates a complex set of presentations achieved from the different agents' actions. Portion of good performing agent group (i.e. an general form) is shared amongst the different agents via a specific form(Q_i) [9]. Such specific forms embrace the limited details about the environment. Such strategies are incorporated to improve the sum of the partial rewards received using satisfactory cooperation prototype. The action plans or forms are created by the way of multi-agent Q-learning algorithm by constructing the agents to travel for the most excellent form Q* and accumulating the rewards. When forms Q1, …, Qx are incorporated, it is possible to construct new forms that is General Form (GF = {GF_1, …, GF_x}), in which GF_i denotes the outstanding reinforcement received by agent i all through the knowledge mode [10]. Algorithm 1 expresses get_form algorithm that splits the agents' knowledge. The forms are designed by the Q-learning used for all prototypes. Outstanding reinforcements are liable for GF which compiles all outstanding rewards. It will be shared by the way of the added agents [11, 12]. Transforming incomplete rewards as GF is considered for outstanding reinforcements to achieve the cooperation between the agents. A status utility gives the outstanding form amongst the opening states and closing state for a known form which approximates GF with the outstanding reinforcements. The status utility is calculated by summation of steps the agent needed to get to destination at the closing state and the sum of the received status in the forms amongst each opening and the closing state [13].

Algorithm1: Cooperative Multi-agent Reinforcement Learning Model

```
Algorithm get_form
1. Initialization Q_i(s, a) and GF_i(s, a)
2. for each agents i ϵ I;
3. agents collaborate till the closing state is estab-
lish;
period ← period +1
4.   Determine the rewards by equation;
 Q(s, a)← Q(s, a) +α (r + γ Q(s', a') − Q(s, a))
5. Fcooperate (period, tech, s, a, i);
6.  Q_i←GF that is Q_i of agent i ϵ I is customized by means
of GF_i.
```

The Fcooperate utility selects a coordination method. period, tech, s, a, I are the factors, in which period is current iteration, cooperation tech is {grp, dyna, gol}, s and a is state and action chosen likewise;

3.1 Cooperation Models

Various cooperation methods for cooperative reinforcement learning are proposed:

(i) *Grp model* – reinforcements are disseminated in a series of periods.
(ii) *Dyna model* – reinforcements are distributed in each action.

Algorithm 2 Cooperation model

```
Fcooperate (period, tech,s,a,i) /*cooperation among
agents as four cases*/
q : count of period
1. Switch between cases
2. In case of Grp method
         if period mod q = 0 then
         get_Policy(Q_i, Q*,GF_i);
3. In case of Dyna method
          r ← ∑_{j=1}^{x} Qj(s,a);
          Q_i(s,a)← r;
          get_Policy(Q_i, Q*,GF_i);
```

Algorithm 3 get_Policy

```
Function get_Policy(Qi,Q*,GFi) /*find out universal agent
policy */
1. for loop for each agent i ϵ I
2. for loop for each state s ϵ S
3.  if status(Qi, s) ≤ status(Q*,s) then
        GFi(s,a) ← Qi(s,a);
 4. end for loop
```

Grp Model: During the learning period each agents collect expertise depend rewards received from their actions. At the end of the period (step q), every agent gives cost of Qj to GF. The usefulness of another agents for given state is enhanced when reward value is appropriate. And these expertise base reinforcements will afterward supplied to the agents. Agent will carry on to make use of its rewards with the objective is for congregating latest values [11–13].

Dyna Model: The coordination in the dyna method is gained as: each act perceived by agent produces a reinforcement value (+ or −), that is summation of all together expertise depends rewards to all agents to action a achieved in state s. Each agent collaborate to achieve more the rewards sum fulfill its own policy [14].

4 Model Design

In practical environment, traffic flows of four signals with eight flow directions are considered for the development. The control coordination between the intersections can be viewed as a Markov process, denoted by $\langle S, R \rangle$ where S represents the state of the intersection, A stands for the action for traffic control and R indicates the return attained by the control agent [15].

4.1 States of System

Instantaneous traffic states are received by each agent. To present state of the road, it returns traffic control decision. Essential data such as number of vehicles newly arriving and number of vehicles currently stopped at signal are used to reflect the state of road traffic [14, 15].

Number of vehicles newly arriving = $X_{max} = x_1, x_2, x_3, x_4 = 10$
Number of vehicles currently stopped at junction J = $I_{max} = i_1, i_2, i_3, i_4 = 20$
State of the system become **Input** as (x_i, i_i).

Here, it can get together 200 possible states by combining maximum 10 arriving vehicle and maximum 20 vehicles stopped at signal (10 * 20 = 200).

4.2 Actions of System

Each policy denotes the learning agent activities at a given time in case of reinforcement learning framework. Rewards are obtained by mapping the scene to the action in reinforcement learning. It affects not only to the next scene but also to direct rewards due to which all successive rewards will be affected [15, 16]. In the study, traffic lights control actions can be categorized to 3 types: no change in signal duration, increasing signal duration, reducing signal duration.

Value	Action
1	No change in signal duration
2	Increase in signal duration
3	Reduce the signal duration

Action set for signal agent 1 is A1 = {1, 2, 3}, action set for signal agent 2 is A2 = {1, 2, 3} and action set for signal agent 3 is A3 = {1, 2, 3}.

Each of them is for one of the following actual traffic scenarios.

The strategy of no change in signal duration is used in the case of the normal traffic flow when the lights control rules do not change [16–18]. The strategy increasing the signal duration is mostly used in the case that in one route is regular and the other route traffic flow is stopped. Two cases are possible i.e. to increase the signal duration to extend the traffic flow and to decrease signal duration when traffic flow on one route is less as compared to other route. Waiting time of other route is reduced as decreased in signal light so that vehicles pass the junction faster.

4.3 Definitions of Reward and Return

Reward function in reinforcement learning describes the target of the problem. The apparent state of the environment is mapped to a value, reinforcement, defining internal needs of the state [18].

In the work, agent makes signal control decisions under diverse traffic circumstances and returns an action sequence, so that by the actions the road traffic jamming display is the least amount. To be additional, the model provides a best traffic synchronization mode in a particular traffic state. Here, we use traffic value display to estimate the traffic flows as

Reward is calculated in the system as given below:

Assume current state i = (x_i, i_i) and next state j = (x_j, i_j). i.e. current state i → next state j

Case 1: $[x_i, i_i] \rightarrow [x_i, i_{i-1}]$ i.e. $[X_{max} = 10, I_{max} = 20] \rightarrow [X_{max} = 10, I_{max} = 19]$
That means: one vehicle from currently stopped vehicle is passing the junction

Case 2: $[x_i, i_i] \rightarrow [x_{i+1}, i_{i-1}]$ i.e. $[X_{max} = 9, I_{max} = 20] \rightarrow [X_{max} = 10, I_{max} = 19]$
That means: one newly arrived vehicle at junction & one vehicle is passing junction

Case 3: $[x_i, i_i] \rightarrow [x_i, i_{i-3}]$ i.e. $[X_{max} = 10, I_{max} = 20] \rightarrow [X_{max} = 10, I_{max} = 17]$
That means: More than one stopped vehicles are passing the junction

Case 4: $[x_i, 0] \rightarrow [x_{i+1}, 0]$ i.e. $[X_{max} = 2, I_{max} = 0] \rightarrow [X_{max} = 3, I_{max} = 0]$

That means: new one new vehicle is arriving and no stopped vehicle at the junction. Depending on above state transitions from current state to next state, reward is calculated as

$$\begin{aligned} \text{Reward is } r_p(i, p, j) &= 1 && \text{if } x_1' = x_1 + 1 \ldots\ldots\ldots \text{Case 4} \\ &= 2 && \text{if } i_1' = i_1 - 1 \ldots\ldots\ldots \text{Case 1} \\ &= 3 && \text{if } i_1' = i_1 - 3 \ldots\ldots\ldots \text{Case 2 \& 3} \\ &= 0 && \text{otherwise} \end{aligned}$$

5 Experimental Results

The study learn a controller with learning rate = 0.5, discount rate = 0.9, and λ = 0.6. During learning process, cost was updated 1000 with 6000 episodes.

The grp method appears to be extremely strong converging very fast to an optimal action form Q*. Rewards obtained by the agents are produced in series of pre identified stages. They gather reasonable reward values that cause a good convergence. In the grp method the global policy converges to a best action strategy as there is an intermission of series necessary to gather good reinforcements. The general form of the dyna method is capable to assemble good reward values in small knowledge series. It is observed that after some series, the performance of global strategy reduces. This takes place since the states neighboring to the final state begin in the direction of more superior reward values giving to a restricted maximum. It will no more stay at the other states so it punishes the agent. In the dyna method as the reinforcement learning algorithm renews learning values, actions with higher gathered reinforcements are chosen through top likelihood than acts with small gathered reinforcements.

Figures 1 and 2 respectively shows that delay time vs number of state given by simple Q learning (without cooperation) and grp and dyna methods (with cooperation). Delay time obtained by cooperative methods i.e. grp and dyn methods is much less than that of without cooperation method i.e. simple Q learning for agent 1 in multi-agent scenario.

Figures 3 and 4 respectively shows that delay time vs number of state given by simple Q learning (without cooperation) and grp and dyna methods (with cooperation). Delay time obtained by cooperative methods i.e. grp and dyna methods is much less than that of without cooperation method i.e. simple Q learning for agent 2 in multi-agent scenario.

Figures 5 and 6 respectively shows that delay time vs number of state given by simple Q learning (without cooperation) and grp and dyna methods (with cooperation). Delay time duration obtained by cooperative methods i.e. grp and dyna methods is much less than that of without cooperation method i.e. simple Q learning for agent 2 in multi-agent scenario.

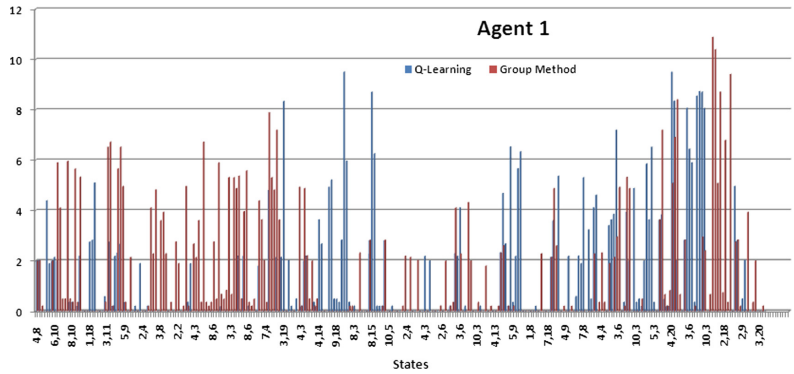

Fig. 1. States vs delay time for agent 1 by Q-learning & grp method

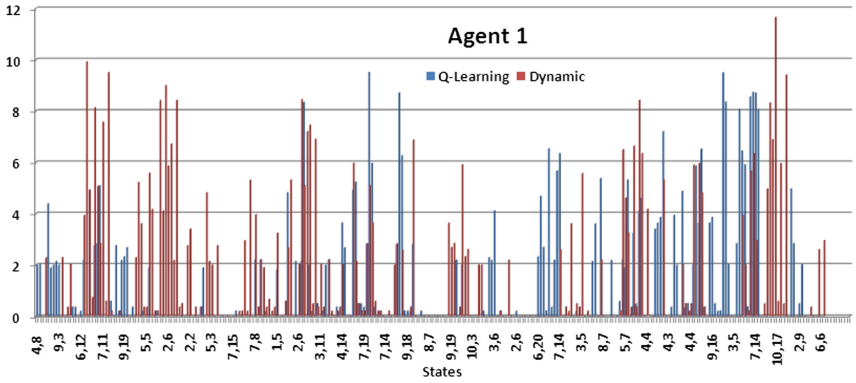

Fig. 2. States vs delay time for agent 1 by Q-learning & dyna method

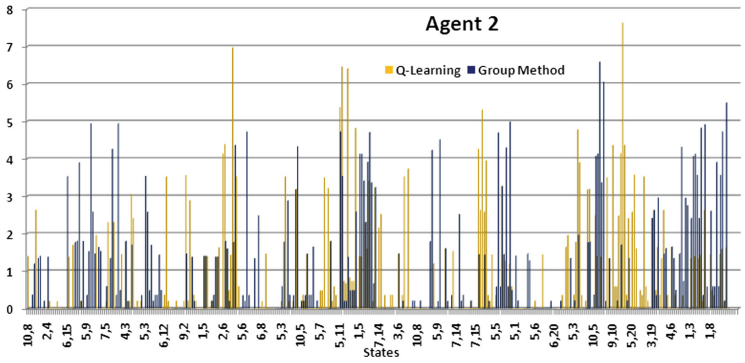

Fig. 3. States vs delay time for agent 2 by Q-learning & grp method

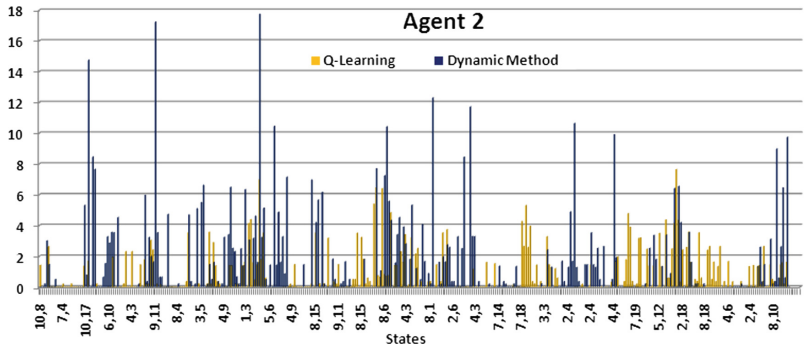

Fig. 4. States vs delay time for agent 2 by Q-learning & dyna method

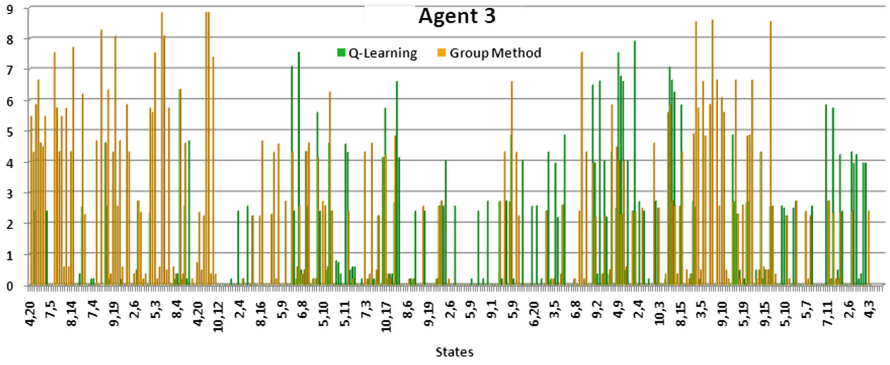

Fig. 5. States vs delay time for agent 3 by Q-learning & grp method

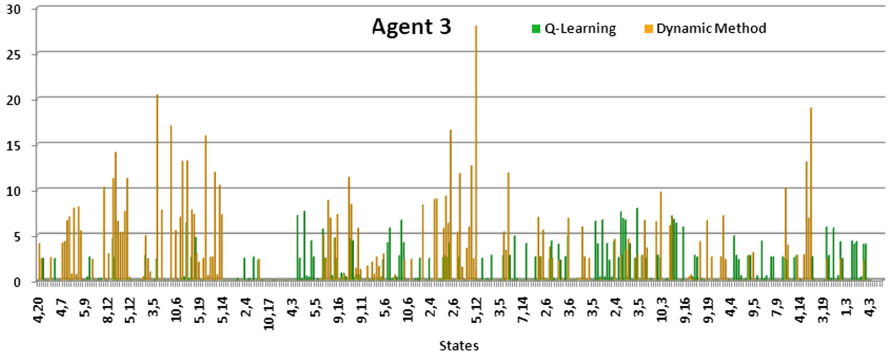

Fig. 6. States vs profit for agent 3 by Q-learning & dyna method

6 Conclusion

Traffic control system is so complicated and dynamic in nature. It is impossible to manage traffic jam and sudden traffic accidents for Q learning model without cooperation with predefined strategy. The demand is getting more and more urgent for combining timely and intelligent traffic control policy with real-time road traffic. Reinforcement learning collects information by keeping communication with situation. Although it usually needs a long duration to complete learning, it has good learning ability to complex system, enabling it to handle unknown complex states well. The application of reinforcement learning in traffic management area is gradually receiving more and more concerns. The paper proposed a cooperative multi-agent reinforcement learning algorithm (CMRLA) for traffic control optimization. The actual continuous traffic states are discretized for the purpose of simplification. Actions for traffic control are designed and rewards are defined to return by mean of traffic cost which combines with multiple traffic capacity indicators.

References

1. Zhu, F., Ning, J., Ren, Y., Peng, J.: Optimization of image processing in video-based traffic monitoring. Elektronika ir Elektrotechnika **18**(8), 91–96 (2012)
2. de Schutter, B.: Optimal traffic light control for a single intersection. In: Proceedings of the American Control Conference (ACC 1999), vol. 3, pp. 2195–2199, June 1999
3. Findler, N., Stapp, J.: A distributed approach to optimized control of street traffic signals. J. Transp. Eng. **118**(1), 99–110 (1992)
4. Vidhate, D.A., Kulkarni, P.: Innovative approach towards cooperation models for multi-agent reinforcement learning (CMMARL). In: Unal, A., Nayak, M., Mishra, D.K., Singh, D., Joshi, A. (eds.) SmartCom 2016. CCIS, vol. 628, pp. 468–478. Springer, Singapore (2016). https://doi.org/10.1007/978-981-10-3433-6_56
5. Baskar, L.D., Hellendoorn, H.: Traffic management for automated highway systems using model-based control. IEEE Trans. Intell. Transp. Syst. **3**(2), 838–847 (2012)
6. Vidhate, D.A., Kulkarni, P.: New approach for advanced cooperative learning algorithms using RL methods (ACLA). In: Proceedings of the Third International Symposium on Computer Vision and the Internet, VisionNet 2016, pp. 12–20. ACM DL (2016)
7. Mase, K., Yamamoto, H.: Advanced traffic control methods for network management. IEEE Mag. **28**(10), 82–88 (1990)
8. Vidhate, D.A., Kulkarni, P.: Performance enhancement of cooperative learning algorithms by improved decision making for context based application. In: International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), pp. 246–252. IEEE Xplorer (2016)
9. Baskar, L.D., de Schutter, B., Hellendoorn, J., Papp, Z.: Traffic control and intelligent vehicle highway systems: a survey. IET Intell. Transp. Syst. **5**(1), 38–52 (2011)
10. Choi, W., Yoon, H., Kim, K., Chung, I., Lee, S.: A traffic light controlling FLC considering the traffic congestion. In: Pal, N.R., Sugeno, M. (eds.) AFSS 2002. LNCS (LNAI), vol. 2275, pp. 69–75. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45631-7_10
11. Vidhate, D.A., Kulkarni, P.: Enhancement in decision making with improved performance by multiagent learning algorithms. IOSR J. Comput. Eng. **1**(18), 18–25 (2016)

12. Wiering, M.A.: Multi-agent reinforcement learning for traffic light control. In: 17th International Conference on Machine Learning (ICML 2000), pp. 1151–1158 (2000)
13. Vidhate, D.A., Kulkarni, P.: Multilevel relationship algorithm for association rule mining used for cooperative learning. Int. J. Comput. Appl. **86**(4), 20–27 (2014)
14. Zegeye, S., de Schutter, B., Hellendoorn, J., Breunesse, E.A., Hegyi, A.: A predictive traffic controller for sustainable mobility using parameterized control policies. IEEE Trans. Intell. Transp. Syst. **13**(3), 1420–1429 (2012)
15. Vidhate, D.A., Kulkarni, P.: A novel approach to association rule mining using multilevel relationship algorithm for cooperative learning. In: 4th International Conference on Advanced Computing and Communication Technologies, pp. 230–236 (2014)
16. Chin, Y.K., Wei, Y.K., Teo, K.T.K.: Q-learning traffic signal optimization within multiple intersections traffic network. In: Proceedings of the 6th UKSim/AMSS European Symposium on Computer Modeling and Simulation (EMS 2012), pp. 343–348, November 2012
17. Vidhate, D.A., Kulkarni, P.: To improve association rule mining using new technique: multilevel relationship algorithm towards cooperative learning. In: International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA), pp. 241–246. IEEE (2014)
18. Chin, Y.K., Lee, L.K., Bolong, N., Yang, S.S., Teo, K.T.K.: Exploring Q-learning optimization in traffic signal timing plan management. In: Proceedings of the 3rd International Conference on Computational Intelligence, Communication Systems and Networks (CICSyN 2011), pp. 269–274, July 2011