Research Issues in Reinforcement Learning for IoT/CPS Domains

XXX YYYY, and XXX ZZZZ

***Abstract*—Reinforcement learning (RL) algorithms are the platform for connecting intelligent Internet-of-Things (IoT) devices. The data provided by these intelligent devices can offer more insights about the overall futuristic functionalities in terms of usage and preferences. Machine learning algorithms have been widely used for predictions, with several optimization algorithms using reinforcement learning techniques to improve the performance and interaction of devices within a dynamic IoT ecosystem. We present a comprehensive review of state-of-the-art RL techniques that have been applied to solve trending problems in the IoT and cyber-physical systems domain. Furthermore, we give an in-depth analysis into such problems and propose alternative approaches that may yield better outcomes.**

***Index Terms*—Reinforcement learning (RL), Internet-of-Things (IoT), cyber-physical systems (CPS), Markov decision process (MDP), machine learning (ML).**

I. Introduction

REINFORCEMENT learning algorithms are the platform for connecting intelligent Internet-of-Things (IoT) devices. Over the years, there has been several applications of machine learning (ML) in IoTs and Cyber-physical systems (CPS). Nest learning thermostats are good examples of how IoT devices leverage data patterns to predict the preferred temperature in a room during a particular time of day. The prediction of the room temperature can also be on an aggregated neighborhood level, where energy loads can be remotely shifted by the power utility in homes operating Nest devices. Another practical application is the Amazon personal assistant that has the capability of learning voice patterns, the Jaguar's Land Monitoring system, which depends on a complex software that allows the automobile to observe, predict, monitor and notify the car's passengers to assist the driver automatically delegate his tasks and minimize the burden of driving.

Optimization is a very old field with interesting algorithms that has been used to solve simple to complex problems in various fields. Many optimization algorithms use RL techniques optimize the behaviour of devices in the IoT ecosystem. For instance, in Intelligent Transport Systems (ITS), where cars act smartly, optimality in the interaction of the cars with the environment is highly required. Several CPS applications like the Industrial Internet of Things (IIoT), smart grid, and ultimately smart cities, will allow intelligent machine-type interactions. These interactions may often will require robust algorithms that evolve and are adaptable, in order for the IoT devices to be able to achieve desired objectives given some set of operational parameters.

The contribution of our work is four-fold, which are listed as follows.

1) We provide a exhaustive review of several RL algorithms, and highlight key aspects that will drive future IoT networks
2) We carried out a detailed comparative study of various RL mechanisms. We also perform simulations on some techniques considering some IoT scenarios.
3) We present detailed RL applications and use cases within the CPS (Intelligent transportation systems, smart grid, smart homes, smart health-care, and smart environment) to demonstrate how different techniques presented in the paper fuse to provide desirable objectives.
4) We also present a variety of open research challenges and suggest possible future trends for building intelligence in IoT, with regard to the latest development in the field.

The remainder of this work is organized as follows. In Section

II. Reinforcement Learning in IoT

Reinforcement Learning (RL) is learning that involves mapping situations to actions with an objective of maximizing a numerical reward. Actions are made by an agent, which have the ability to sense the state of the dynamic environment and consequently take actions that influence its environment. RL facilitates sequential decision making under uncertainty, thereby making it a useful tool in prediction of non-linear phenomenon [2]. Several works have been published in the area of RL in IoT and CPS, however, there has not been any detailed review that covers their applicability in IoT and the underlying research issues in this field.

Over the years, many perceived difficult subproblems have began to receive research attention.

A. Q-Learning

Q-learning is also known as an off-policy temporal difference (TD) learning algorithm, which allows the agent to learn about an optimal policy using an exploratory policy [3], [4], [5]. The Q-learning algorithm can be simply defined by

$$Q_{t+1}(s_t, a_t) := Q_t(s_t, a_t) + \alpha\Big[r_{t+1} + \gamma \max_a Q_t(s_{t+1}, a) - Q_t(s_t, a_t)\Big]. \quad (1)$$

Manuscript received March 14, 2019; revised July X, 2019.

B. W-Learning

Unlike Q-learning which is a single-agent, single-policy learning technique, W-learning is a multi-agent, multi-policy learning technique that has been used on non-cooperating agents [9]. The W-learning algorithm is given by

$$W_i(s) := (1-\alpha)W_i(s) + \alpha\Big[Q_i(s,a_i) - (r_i + \gamma \max_a Q(s', a_i'))\Big]. \tag{2}$$

C. SARSA

State-action-reward-state-action (SARSA) is an On-policy TD learning algorithm, which the agent learns an action-value function instead of a state-value function. SARSA always converges to an optimal policy so long as all state-action pairs are visited an infinite number of times.

$$Q_{t+1}(s_t,a_t) := Q_t(s_t,a_t) + \alpha\Big[r_{t+1} + \gamma Q_t(s_{t+1},a_{t+1}) - Q_t(s_t,a_t)\Big]. \tag{3}$$

D. Deep Q Net (DQN)

E. Deep Deterministic Policy Gradients (DDPG)

F. Normalized Advantage Functions (NAF)

G. Asynchronous Advantage Actor-Critic (A3C)

TABLE I
SUMMARY OF WORKS IN REINFORCEMENT LEARNING.

Reinforcement learning techniques	Details	Limitations	Potential contributions
RL-based mapping table (RLMT) and RL-based resource allocation	These algorithm was introduced Gai *et al.* in [8] to handle cost mapping tables creation and for optimal resource allocation in IoT content-centric services.	The work was limited to the use QoE in examining service level, which allowed a dynamic resource allocation and avoided a fixed task table.	To apply RL-based approach on several other KPIs that may improve IoT services.
Deep-Q-learning	This variant of Q-learning was introduced by Zhu *et al.* in [6] to maximize system throughput by applying appropriate scheduling strategy for cognitive radio-based IoT networks.	This work performed poorly when compared with the strategy iteration algorithm, though with a reduced computational complexity.	To adapt the concept of cognition to our proposed IoT network and with the consideration of several network dynamics.
Deep Reinforcement Learning (DRL)	A semi-supervised DRL model that suits IoT and smart city applications was introduced in Mohammadi *et al.* in [14] for performance improvement and accuracy in the learning agent.	The work employed an indoor localization based on the Bluetooth low energy signal strength, and limited its findings to a single floor of a building.	To consider multi-agent environment and conduct some outdoor experimentation.
RL using a Markov-based analytical model	A Markov-based analytical model was integrated with a RL process in Conti *et al.* [17] to optimize the server activation policy, where optimal control of an energy storage system in a green fog-computing node is needed to improve the system performance, hereby allowing the system to bear high job arrivals even at low-power generation periods.	The fog-computing nodes used in this work is fixed and has a large energy source, which fails to depict the resource-constrained nature of fog devices.	We will consider a multi-tier fog architecture with lots of heterogeneity, having static and mobile fog nodes which may or may not be power-constrained.
Q-learning	Wen *et al.* in [4] formulated an automated energy management system (EMS) rescheduling problem as a reinforcement learning (RL) problem. Simulations were carried out using Q-learning technique on a specific scenario with good results.	The paper considers the EMS to act as an agent for energy users. This approach is not practical in a real IoT scenario, where nodes are mobile with dynamic system requirements.	We will consider a decentralized agent-based system to support for a realistic IoT scenario.
Evolutionary strategies and RL	A RL approach that made use of evolution strategies for real-time task assignment among fog servers was introduced by Mai *et al.* in [18] to minimize the total computation latency during a long-term period.	The paper claimed that the proposed model is scalable when the number of IoT devices increases, however, the approach proposed failed to examine real-world IoT-scaled scenario.	We aim to apply a variant of these techniques to realistic IoT scenarios.
Q-learning-based duty cycle control	This RL-based duty cycle control technique was introduced by Li *et al.* in [20] to provide improved performance and reliable M2M communication for IoT applications.	The performance evaluation of the proposed Q-learning based duty cycle control only considered a two-hop cluster tree network.	The proposed approach can be enhanced to capture a larger network size, and DP approach employed may not be suitable.
Dynamic Programming based duty cycle control	This technique used by Li *et al.* in [21] to provide an optimal solution to an inventory control problem. The focus was on optimizing the duty cycle by jointly considering energy efficiency, end-to-end delay and reliability of the network.	A two-hop cluster tree network model was used in the work, and the problem was evaluated using the DP approach	The proposed approach can be enhanced to capture a larger network size, and the DP approach used may not be suitable for a realistic IoT scenario.
Predictive and Resilient Q-learning	A variant of Q-learning algorithm employed by Grammatopoulou *et al.* in [22], which considers historical data about irregular operations such as faults and attacks by malicious agents in an IoT network (a smart water supply system).	xxxx	yyyy.

III. Practical issues that RL could be applied to IoT (thoughts when reading your paper)

To optimize based on the agent's ability learn different policies that are based on the following:

- Power level (battery/energy per time of the relay/agent): Energy is very important in IoT networks as most devices are power-constrained. An agent within the network should be able to make decisions based on its local policies as well as the remote policies from neighbouring agents within the network.
- Communication outage (based on the channel conditions/state of the environment/single hop): Every physical link, which is often wireless has unique channel conditions, and as such, the agent should be able to learn the optimal route.
- Cooperative capability (malicious agents/trusted nodes): In every network, there exist possibilities of having malicious agents. Overtime, the agent should learn how to avoid malicious agents.
- Storage capability: In a distributed and heterogenous network where neighbour nodes/agents have different storage capacity. An RL-based technique could be used by the agent to learn. This will in turn minimize packet loss within the network.
- Mobility patterns/metrics of agents and that of its neighbours: Considering a realistic IoT network where agents can be static or mobile, the node degree of agents in the network will vary with time, and as such it will be necessary for an agent to consider the dynamics of neighbouring agents in the network.
- Estimated transmission delay: The position of neighbours with the network has an important role to play in determining the delay. An agent should be able to learn and follow a policy that best minimizes network latency.
- Noisy and dynamic environment:
- The issue of full observability and partial observability MDPs: POMDPs are known to be intractable only for small problems.

IV. Challenges we may face when implementing the above issues raised

We know that devices/agents in an IoT network are computationally-constrained, it will be proper to deploy lightweight RL-based techniques (or any other machine learning approach) that will improve the efficiency of the network. Furthermore, we may have to experiment on a very dynamic environment considering factors that are equally stochastic, and ultimately allow for strict quality-of-service requirements.

V. Application of Reinforcement Learning in CPS

A. Smart Grid

B. Intelligent Transport System

C. Smart Cities

VI. Conclusion

References

[1] S. Earley, "Analytics, Machine Learning, and the Internet of Things," *IT Professional,* vol. 17, no. 1, pp. 10-13, Jan.-Feb. 2015.

[2] D. Zhang, X. Han and C. Deng,"Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE Journal of Power and Energy Systems,* vol. 4, no. 3, pp. 362-370, September 2018.

[3] C.J.C.H. Watkins and P. Dayan, Machine Learning (1992), Kluwer Academic Publishers, vol. 8: 279.

[4] Z. Wen, D. O'Neill and H. Maei, "Optimal Demand Response Using Device-Based Reinforcement Learning," *IEEE Transactions on Smart Grid,* vol. 6, no. 5, pp. 2312-2324, Sept. 2015.

[5] R. S. Sutton, and A. G. Barto, *Reinforcement learning - an introduction.* Cambridge, MA: MIT Press, 1998.

[6] J. Zhu, Y. Song, D. Jiang and H. Song, "A New Deep-Q-Learning-Based Transmission Scheduling Mechanism for the Cognitive Internet of Things," *IEEE Internet of Things Journal,* vol. 5, no. 4, pp. 2375-2385, Aug. 2018.

[7] J. Zhu, Z. Peng and F. Li, "A transmission and scheduling scheme based on W-learning algorithm in wireless networks," *2013 8th International Conference on Communications and Networking in China (CHINACOM),* Guilin, 2013, pp. 85-90.

[8] K. Gai, M. Qiu, "Optimal resource allocation using reinforcement learning for IoT content-centric services," *Applied Soft Computing,* vol. 70, pp. 12-21, Sept. 2018.

[9] I. Dusparic and V. Cahill, "Distributed W-Learning: Multi-Policy Optimization in Self-Organizing Systems," 2009 Third IEEE International Conference on Self-Adaptive and Self-Organizing Systems, San Francisco, CA, 2009, pp. 20-29.

[10] L. Zhou, A. Swain and A. Ukil, "Q-learning and Dynamic Fuzzy Q-learning Based Intelligent Controllers for Wind Energy Conversion Systems," *2018 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia),* Singapore, 2018, pp. 103-108.

[11] T. Park, N. Abuzainab and W. Saad, "Learning How to Communicate in the Internet of Things: Finite Resources and Heterogeneity," *IEEE Access,* vol. 4, pp. 7063-7073, 2016.

[12] A. Hans and S. Udluft, "Ensembles of Neural Networks for Robust Reinforcement Learning," *2010 Ninth International Conference on Machine Learning and Applications,* Washington, DC, 2010, pp. 401-406.

[13] D. D. Nguyen, H. X. Nguyen and L. B. White, "Reinforcement Learning With Network-Assisted Feedback for Heterogeneous RAT Selection," *IEEE Transactions on Wireless Communications,* vol. 16, no. 9, pp. 6062-6076, Sept. 2017.

[14] M. Mohammadi, A. Al-Fuqaha, M. Guizani and J. Oh, "Semisupervised Deep Reinforcement Learning in Support of IoT and Smart City Services," *IEEE Internet of Things Journal,* vol. 5, no. 2, pp. 624-635, April 2018.

[15] S. Alletto et al., "An Indoor Location-Aware System for an IoT-Based Smart Museum," *IEEE Internet of Things Journal,* vol. 3, no. 2, pp. 244-253, April 2016. doi: 10.1109/JIOT.2015.2506258

[16] K. Kolomvatsos, and C. Anagnostopoulos, "Reinforcement Learning for Predictive Analytics in Smart Cities," *Informatics,* vol. 3, no. 16, June 2017.

[17] S. Conti, G. Faraci, R. Nicolosi, S. A. Rizzo and G. Schembra, "Battery Management in a Green Fog-Computing Node: a Reinforcement-Learning Approach," *IEEE Access,* vol. 5, pp. 21126-21138, 2017.

[18] L. Mai, N.-N. Dao, and M. Park, "Real-time task assignment approach leveraging reinforcement learning with evolution strategies for long-term latency minimization in fog computing," *Sensors,* vol. 18, no. 2830, Aug. 2018.

[19] C. Kwok and D. Fox, "Reinforcement learning for sensing strategies," *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566),* Sendai, 2004, pp. 3158-3163 vol.4.

[20] Y. Li, K. K. Chai, Y. Chen and J. Loo, "Smart duty cycle control with reinforcement learning for machine to machine communications," *2015 IEEE International Conference on Communication Workshop (ICCW),* London, 2015, pp. 1458-1463.

[21] Y. Li, K. K. Chai, Y. Chen and J. Loo, "Optimised delay-energy aware duty cycle control for IEEE 802.15.4 with cumulative acknowledgement," *2014 IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC),* Washington, DC, 2014, pp. 1051-1056.

[22] M. Grammatopoulou, A. Kanellopoulos, and K. G. Vamvoudakis, "A multi-step and resilient predictive Q-learning algorithm for IoT: a case study in water supply networks", *ACM Proceedings of the 8th International Conference on the Internet of Things,* Santa Barbara, California, 2018, pp. 1-8.