A SURVEY ON LEARNING MODELS OF REINFORCEMENT LEARNING

Subhashini.R, subhaaishu15@gmail.com , Ravi Ramya Sree.P, raviramyasree1@gmail.com, and Anitha.S,anithaselvaraj83@gmail.com

***Abstract*—Q-learning is a frequently used *reinforcement learning* method, *as* well as an off-policy strategy for the overall performance since its inception many articles, have described q- learning role in reinforcement learning. Initial Q-learning algorithms were ineffective in a lot of reasons and could only be used to solve a restricted number of issues it has also been established that this rather powerful algorithm learns inappropriately and overestimates the action values on occasion resulting in lower overall performance. Several q-learning variants such as deep q-learning which incorporates basic q *learning with deep neural networks had* also recently been formed as a result of the recent breakthrough in machine learning. As a result of general improvements in machine learning have been encountered and applied comprehensively in recent years. Finally, in this work, we look at the many forms of q-learning algorithms.**

Keywords:, reinforcement learning, Mdp ,q-learning, machine learning

I INTRODUCTION

One of the main aspects of reinforcement learning is trial and error. Machine learning is a notion in which a computer program can adapt and learn new data without the need for human intervention. Machine Learning is specialized, not broad, in that it enables a machine to make predictions or make decisions based on data about a specific situation. In the production of games nowadays, a mix of Machine Learning and Reinforcement Learning is employed. Agents are trained, and readings are gathered to determine the game's efficiency. Certain regulations are built into the game, and game agents are trained to assist in the resolution of certain issues. Reinforcement Learning is one of the most exciting methods in machine learning, in which optimum actions are learnt in any condition through trial and error. This trial-and-error method aids in issue solving and provides a fresh viewpoint on Reinforcement Learning's performance investigation element. In supervised learning, the algorithm learns from a training dataset and provides predictions, which are then compared to the actual output values in this technique. The algorithm is modified until it's perfect if the predictions aren't accurate. [13]. in this unsupervised learning, the algorithm studies the data instead of determining an output for the input. The algorithm is left unsupervised in order to ascertain the fundamental pattern in the data and learn more about it [9]. Reinforcement learning is a type of machine learning that allows the system to learn from its failures. You provide the machine with a given circumstance in which to carry out a series of tasks. It will now learn through trial and error. A prototypical RL system consists of two major parts: an agent and an environment. The agent symbolizes the RL algorithm, whereas the environment refers to the object on which the agent is functioning.

II. LITERATURE REVIEW

YANGYANG GE et al denotes the Reinforcement learning (RL) seeks to solve a problem by taking behaviours that maximise the long-term reward. The agent of reinforcement learning may learn the best policy by interacting with the dynamic environment through trial and error given a specified state-action combination [1]. Despite the fact that reinforcement learning algorithms have been researched for a long time, the majority of them were originally built for toy contexts [17].

Zhou Ke et al [18] depict about the data representation (board and rule representation), search, evaluation, move generation, knowledge base, and other aspects are all included in a perfect computer game system. The evaluation is a difficult and crucial aspect of them all. The feature parameters have a big influence on the evaluation function's design. The parameters have a significant impact on the efficiency and precision of the evaluation. As a result, after a given initial value, we must adjust them repeatedly to achieve the ideal state.[18]

XueJinlin et al [19] talks about The Action Selection Network (ASN) acts as an actor in the neural network with reinforcement learning, while the Action Evaluation Network (AEN) criticises the ASN's actions. The two networks' outputs feed into the Stochastic Action Modifier (SAM), which

investigates effective actions by assigning a stochastic deviation to the ASN's actions.[19]

Lucileide M.D.DA Silva et al [20] symbolises that , FPGAs are hardware platforms that are well suited for the deployment of software algorithms. Based on the theoretical foundation presented, it is possible to conclude that FPGA devices have a faster execution time than their software counterparts. the main reason for its use as a development platform for the Q-learning Learning Technique.[20]

Snehasis Mukhopadhyay et al[21] represents the Markov Decision Process (MDP) hastwo learning automata are used to solve uncertain zero-sum game problems. Some of these methods are based on adaptive Shapley Recursion, and minimax Temporal Difference. A novel algorithm based on Heterogeneous Games of Learning Automata (HEGLA) is also presented.[21]

III. RELATED WORK

AUTHORS	YEAR	METHODOLOGY	DATASET
XueJinlin et al	2010	Reinforcement learning	Engine idle
Martin van Otterlo, Marco Wiering	2012	Markov decision process	MDP using Bellman's equation
Hasselt et Al	2015	Reinforcement learning	Deep Reinforcement Learning with Double Q-learning
Thomas Hubert et Al	2017	Implementation of RL	Self-play chess game
ErsinSelvi Et al	2018	Markov decision process	Cognitive radar
JitsSchilperoort et AL	2018	Q-learning	PLAY PAC-XON
Massimiliano Patacchiola, Amos Storkey	2020	Supervised Learning	Relational reasoning
Ali Jaber Almalki	2019	Reinforcement learning	Play snake game

IV LEARNING MODELS OF REINFORCEMENT LEARNING

A.Markov Decision Process:

Markov Decision Process (MDP) is a method for making decisions based on observations in a sequential manner and being rewarded for attaining specific states. The learning agent, the environment, a policy, a reward function, and a value function are the five main components of the dynamic environment of the game to be solved with reinforcement learning [11]. The policy is a mapping from states to actions that result in a series of state-action pairings that describe a learning agent's overall behaviour. The policy determines and executes the action (at) at time t based on a certain state (st). The learning agent goes through the training stage as the procedure begins. The learning agent does not have any knowledge about its surroundings at first, therefore its policy decides for it.[15]

Important terms in MDP:

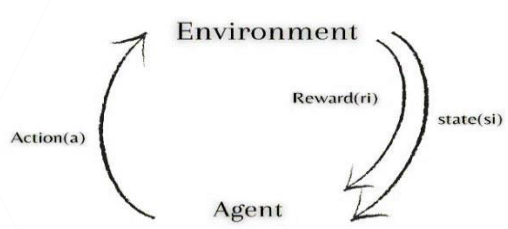
Set of actions- A

Set of states -S

Reward- R

Policy- n

Value- V



MDPs are used to simulate autonomous agent planning in an unpredictable environment. MDPs are popular in two artificial intelligence subfields Probabilistic planning and reinforcement learning are two types of reinforcement learning. The research on probabilistic planning focuses on creating computationally efficient ways to solving

MDPs. assumption that you have comprehensive understanding of the MDP .Reinforcement learning, on the other hand, is a more complex task. [4] discussed that the activity related status data will be communicated consistently and shared among drivers through VANETs keeping in mind the end goal to enhance driving security and solace. Along these lines, Vehicular specially appointed systems (VANETs) require safeguarding and secure information correspondences. [6] discussed because of various appealing focal points, agreeable correspondences have been broadly viewed as one of the promising systems to enhance throughput and scope execution in remote interchanges.

A situation in which the agent has no prior information of the MDP and must engage with others to learn from their experiences. and gaining information by exploring with its surroundings discussing ways to make it behave better [11].As time goes, the agent has a greater understanding of its surrounding and establishes a policy derived from past rewards.An agent learns the best behaviour in a given environment through rewards and punishment (negative reward) in Reinforcement Learning, a machine learning approach. Given his current situation, the agent must discover the best course of action to take. The problem is known as the Markov Decision Procedure when this process is performed several times (MDP). [2] proposed a secure hash message authentication code. A secure hash message authentication code to avoid certificate revocation list checking is proposed for vehicular ad hoc networks (VANETs). [8] discussed that Helpful correspondence is developing as a standout amongst the most encouraging procedures in remote systems by reason of giving spatial differing qualities pick up. The transfer hub (RN) assumes a key part in agreeable correspondences, and RN choice may generously influence the execution pick up in a system with helpful media get to control (MAC).

MDP is a methodology that assists in the choosing of a set of behaviours that maximise reward in a stochastic environment. It is based on the assumption that the Markov condition is fulfilled, which implies that the impact of a single action in a certain state is solely dependent on that state and not on prior ones

MDP Model S: State space is a collection of all conceivable states.[12]

B. Q learning:

Q-learning, which started out as an incremental algorithm for estimating the optimal decision strategy in an infinite-horizon decision issue, has evolved into a broad category of reinforcement learning techniques commonly employed in statistics and artificial intelligence. Q-learning employs an off-policy control that isolates the deferral policy from the learning policy and uses the Bellman optimal equations and the e-greed policy to update the action selection. [1] Because Q-learning features simple Q-functions compared to other reinforcement learning algorithms, it has formed the cornerstone for many additional reinforcement learning algorithms.

Types of Q-learning Algorithms:

1) Algorithms of Single agent type:

Basic Q learning:

To decouple the acting policy from the learning policy, Q-learning employs an off-policy technique.

As a consequence, even if the action chosen in the following state was mediocre, the information was not incorporated in the updating of the current state's Q-function, resulting in the illusion that it was a bad decision.[3] Q-learning, on the other hand, overcomes the problem since it employs off-policy. The following is the Q-value equation:

$$Q(s, a) \leftarrow Q(s, a) + \alpha[R + \gamma \max Q(s0 , a\ 0) - Q(s, a)]$$

Deep Q learning:

Deep Q-learning, created by Google Deep Mind, combines Convolution Neural Networks (CNN) and basic Q-learning. Apart from the value approximation using a CNN, deep Q-learning incorporates two techniques. One is the target Q technique, while the other is an experience replay.

Hierarchical Learning:

When the state-action space of Q-learning expands, complications develop. Hierarchical Q-Learning is designed to address these issues[5]. The abstract action, which splits the agent's activity into a higher and lower level, is the main concept of hierarchical Q-learning.

Double Q learning:

Q-learning does not function well in a stochastic setting [7], hence double Q-learning was created to address this issue. After a given amount of time, traditional Q-learning does not seek for a new optimal value, but instead picks the highest value from current values. Double Q-learning splits the Q-learning valuation function, which defines the measure to stop the value from deviating in the Q-learning process. [10] discussed about diabetic retinopathy from retinal pictures utilizing cooperation and information on state of the art sign dealing with and picture preparing. The Pre-Processing stage remedies the lopsided lighting in fundus pictures and furthermore kills the fight in the picture.

In addition, in a single-agent setting, there are a number of algorithms that use Q-learning. Incremental multistep Q-learning, asynchronous stochastic approximation Q-learning, and Bayesian Q-learning are common examples.

2) Algorithms of multiagent types:

Modular Q learning:

Modular Q-learning was created to address the inefficiencies of basic Q-learning in multi-agent systems. Modular Q-learning overcomes the huge state-space difficulty of Q-learning by breaking down a large problem into smaller sub-problems and implementing Q-learning to each one.

Ant Q learning:

An ant system (AS) and Q-learning are combined in Ant Q-learning. AS is an algorithmic portrayal of ants returning to their colony after obtaining food. [9] Unlike traditional Q-learning, ant Q-learning employs many agents to learn. Because agents in ant Q-learning interact with one another, it is possible to successfully identify the value of the reward for a specific activity in a multi-agent environment.

Swarm based Q learning:

When the learning issue is complicated, a standard Q-learning method takes a long time to identify the best solution. Furthermore, in a multi-agent system, the solution is frequently unavailable or takes a long time to find. Particle Swarm Optimization (PSO) is used in swarm-based Q-learning to discover the best solution. With a large solution space, PSO can swiftly identify a globally optimum solution for numerous module functions.

There are also additional algorithms that use Q-learning in a multi-agent context, such as Nash Q learning, that use Q-learning.

V CONCLUSION

In this paper, we started off with Machine Learning and the three different types of it: Supervised learning, Unsupervised learning and Reinforcement learning. Among these we dwelled a little deeper into the final type where we saw that Reinforcement Learning is a method for learning control methods for autonomous entities when there is little or no data. RL is always learning, and as a result, it improves its performance at the job at hand. Then we further moved to see the two models of Reinforcement learning. One of it is Markov decision process which is a sequential, stochastic decision-making method based on the Markov Property. MDPs may be used to make the best decisions for a dynamic system based on its present condition and surroundings. In reinforcement learning applications, this technique is crucial. The second model is Q learning. All versions of Q-learning algorithms, which are a typical algorithm under reinforcement learning, were analysed. We distinguished between single-agent and multi-agent Q-learning techniques and carefully discussed each. Finally, we detailed clearly About Q- learning using all the concepts mentioned in the paper.

REFERENCES

[1] C. J. C. H. Watkins and P. Dayan, ‘‘Q-learning,’’ Mach. Learn., vol. 8, nos. 3–4, pp. 279–292, 1992.

[2] Christo Ananth, M.Danya Priyadharshini, “A Secure Hash Message Authentication Code to avoid Certificate Revocation list Checking in Vehicular Adhoc networks”, International Journal of Applied Engineering Research (IJAER), Volume 10, Special Issue 2, 2015,(1250-1254).

[3] T. G. Dietterich, ‘‘Hierarchical reinforcement learning with the MAXQ value function decomposition,’’ J. Artif. Intell. Res., vol. 13, pp. 227–303, Nov. 2000.

[4] Christo Ananth, Dr.S. Selvakani, K. Vasumathi, “An Efficient Privacy Preservation in Vehicular Communications Using EC-Based Chameleon Hashing”, Journal of Advanced Research in Dynamical and Control Systems, 15-Special Issue, December 2017,pp: 787-792

[5] A. G. Barto and S. Mahadevan, ‘‘Recent advances in hierarchical reinforcement learning,’’ Discrete Event Dyn. Syst., vol. 13, nos. 1–2, pp. 41–77, 2003.

[6] Christo Ananth, Dr. G. Arul Dalton, Dr.S.Selvakani, “An Efficient Cooperative Media Access Control Based Relay Node Selection In Wireless Networks”, International Journal of Pure and Applied Mathematics, Volume 118, No. 5, 2018,(659-668).

[7] H. van Hasselt, A. Guez, and D. Silver, ‘‘Deep reinforcement learning with double Q-learning,’’ 2019, arXiv:1509.06461. [Online]. Available: https://arxiv.org/abs/1509.06461.

[8] Christo Ananth, Joy Winston.J., “SPLITTING ALGORITHM BASED RELAY NODE SELEC-TION IN WIRELESS NETWORKS”, Revista de la Facultad de Agronomía, Volume 34, No. 1, 2018,(162-169)

[9] M. Dorigo and L. M. Gambardella, ‘‘A study of some properties of ant-Q,’’ in Proc. Int. Conf. Parallel Problem Solving Nature, 1996, pp. 656–665

[10] Christo Ananth, D.R. Denslin Brabin, Jenifer Darling Rosita, “A Deep Learning Approach To Evaluation Of Augmented Evidence Of Diabetic Retinopathy”, Turkish Journal of Physiotherapy and Rehabilitation, Volume 32,Issue 3, December 2021,pp. 11813-11817

[11]ErsinSelvi and R. Michael Buehrer and Anthony Martone and Kelly Sherbondy :”On The Use of Markov Decision Processes in cognitive radar: an application to target tracking" InComputer Science IEEE Radar Conference., Apr.2018

[12]Garima Gupta and Rahul Katarya: “ A Study of Recommender Systems using Markov Decision Process” in Proc of the .,(ICICCS 2018)

[13] A. Hammoudeh, “A Concise Introduction to Reinforcement Learning,” 2018 .

[14] Ali Jaber Almalki Pawel Wocjan : “Exploration of Reinforcement Learning to Play Snake Game” In proceedings of the ., International Conf.,on Computational Science and Computational Intelligence (CSCI).,2018

[15]JitsSchilperoort and Ivar Mak and Madalina M. Drugan and Marco A. Wiering: “Learning to Play Pac-Xon with Q-Learning and Two Double Q-Learning Variants” .,published in IEEE(2018)

[16] Beakcheol Jang and Myeonghwi Kim and Gaspard Harerimana and Jong Wook Kim:” Q-Learning Algorithms: A Comprehensive Classification and Applications”., published in IEEE (SEP 27 TH 2019)

[17] YANGYANG GE1,FEI ZHU 1,2, XINGHONG LING1, AND QUAN LIU: “Safe Q-Learning Method Based on Constrained Markov Decision Processes “ published in IEEE Access Volume: 7(Nov 11, 2019)

[18] Zhou Ke, Wong Huan, Wu Ruo-fan, Qi Xin:” An Improved Algorithm Model based on Machine Learning” Published in The 27th Chinese Control and Decision Conference (2015 CCDC).,IEEE EXPLORE(2015)

[19]Xue, Jinlin; Gao, Qiang; Ju, Weiping :” Reinforcement Learning for Engine Idle Speed Control” [IEEE 2010 International Conference on Measuring Technology and Mechatronics Automation]

[20]Lucileide M. D. Da Silva; Matheus F. Torquato; Marcelo A. C. Fernandes:“Parallel Implementation of Reinforcement Learning Q-Learning Technique for FPGA”., published in: IEEE Access (Volume: 7) Pp:2782 – 2798., 13 dec (2018)

[21]SnehasisMukhopadhyay,OmkarTilak,Subir Chakrabarti :” Reinforcement Learning Algorithms for Uncertain,Dynamic, Zero-Sum Games”.,published in 2018 17th IEEE International Conference on Machine Learning and Applications