# Analysis on Reinforcement Learning Technique Based on Q-Learning

Shashikala Mishra[1] Varun Singh[2]

[1] *M.Tech Research Scholar,* [2] *Assistant Professor*

*Rewa Institute of Technology, Rewa (M.P.)*

***Abstract - Artificial intelligence described as: how to agent learns automatically and executed them by understanding. In current years several successful applications of the artificial intelligence had been designed for clustered data base programs which learn to detect deceptive credit card transactions to passes over system, learn user reading preferences to autonomous vehicles to move rapidly on roads. There many advances in the theory and algorithms transformed several derived concepts towards it and this paper contains its survey report and its analysis.***

***Keywords: Reinforcement learning system, Decision making system, Look up table, Reward, Agent, Discount rate.***

## I. INTRODUCTION

Artificial intelligence is innovative concepts of present scenario. This paper introduces reinforcement learning techniques for cell problem in the field of machine learning and its application. So review about different topics and concepts regarding data learning. Thus, the study about query based reinforcement learning, is an effective method of self learning technique through number of episodes.

Many troubles faced by animals, human beings and AI systems can be modeled as sequential decision making in doubtful dynamic environments. For example a complex industrialized system involves optimizing hundreds or even thousands of processes such as catalog, manufacturing drawing, assembly and advertising. These troubles involve decision makers or agents, selecting consecutive action in order to long-term goals. Moreover, uncertainty does come in these domains, both in the effect of actions and valuation of the actual system. In current years, advances in machinery had led to amplified attention in computerized methods for solving these tasks.

In this paper goal is to review how to train the agent and learning using grid world problem. Artificial intelligence application oriented with solution in the field of robotics, cognition, brain theory.

In order theory, biology, cognitive science, computational complexity and control theory.

Reinforcement learning system is an agent, which can a recognize environment, learn to get the most favorable action to attain the objective. In the machine learning number of episode interaction responds towards of rewards or penalty over given environment. Reinforcement learning agent based on when the agent makes any action, system learns from that environment and gives response in the form of signal, which is called reward in terms of positive values. RL is deals with, in case of multi agent learning and their interaction to make best use of as a mathematical return signal.

Decision making system studies the interactions over an agent and its environment. Look up table towards matrix as an environment and also distinguish to take actions or to change. Learning, allows the agent to activate in initially unidentified environments and to become more competent than its original awareness alone strength allow towards reinforcement learning. Thus they try to find possible solution in the form of solstices on the basis of scalar evaluation. During number of episodes it's trained and learns to approach the desired goal [1].

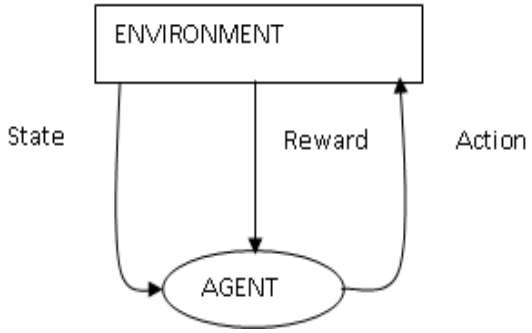



Fig.1.1: Learning Agent and Environment [1].

An agent interaction described through the above fig.1.1 [1] and their interaction reacts towards agent earning activities their performance. Their objective is defined in a way that maximizes potential reward.

RL system [2] having five tuple :{ S, A, π, RF, VF}, defined as, environment defined on the basis of E, set of actions is described in terms of A, performed action over policy π, reward signal R respectively. Policy play roles towards establish relationship between agent and environment.

A policy is responsible for a mapping from a state of the environment action by the agent. They also defined and train agents behavior for taking quick action by agents in the form decision making.

A reward function is a mapping from the state or state action pair of the environment over generated values called reward signals. There is an indication of the desirability of the look up table. Important role of learning system is to diagnosis automatic decision making system and their movement. Discount rate defined using mathematical expression $St=s\gamma^{t\ their}$ range lies between 0 and 1.

Agent learning described subsequences with contenting goal state in the grid world and their movement from current position. [2]. policy also gives the return value that received from environment. State values function defined on the basis of policy equation: $P^{\pi}(s) = E_{\pi}\{s_t Is_t=s\}=E_{\pi}\{\sum_{k=0}^{\infty} r\gamma t\}$,

The action-value function of moves over grid world problem mathematically defined as $Q^{\pi}(s,a)=E_{\pi}\{R_t Is_t=s,a_t=a\}$, this is the predictable come back under policy $\pi$, preliminary from move a in states. Exploits methods based on greedy method is to explore enchanting an action other over policy greedy action. The intention of its examination is to determine other actions that might better than the greedy action [4].

The e-greedy method [5] is one that performs together utilization and examination. With probability1-$\varepsilon$, where $\varepsilon$ is a small positive number, it takes the greedy action, i.e., it exploits and it selects an exploit randomly. Agent movement performed in left right up and down through rewards under policy over environment. Training grid based over environment is represented by the set of states. Agent learning is judgment creator over perceives history and select one from given environment. They estimate the RL for control problem such as grid world problem using q learning reinforcement learning (QRL) techniques. However, these methods address predicting time delayed rewards problem and computes future rewards. Their assessment function is estimated of future recital in terms of commutative rewards.

RL acts as a decision making performer over environment on the basis of trial and error interaction [3]. During learning system agent will take action over grid world in the form of state action and agent moved on next state. Agent revived response in the form of rewards and penalty and also evaluates corresponds values in terms of delayed rewards. After number of episodes agents learns how to reach that entire goal and correct actions performed to produce feedback under less information with the help of supervised learning.

Reinforcement system, defined about the behavior of teaching agent's right mind and acts over grids and train to arrived to entire goal offer a general structure and several methods, to make or to get better behavior, while it movement over grids with large space are not suitable [6][7]. In Reinforcement learning most advantage action performed over grid worlds on the basis of supervised learning. Therefore the learner has to modify its policy though number of episodes [8]. Reinforcement learning trained to the agent's behavior via number of episodes.

So it maximizes commutative reward. It may be analytical concepts method used here for the learning state in the cognition cycle,

Using artificial neural network trained using weighting factors to give accurate prediction on the fixed threshold level [8][9].

Basic concept of self learning is trained data and learns it through discount rate and mapped hidden history. It is to increase commutative rewards, what to do and how to map situations to actions. It is to maximize imitative reward from the environment. The plan of reinforcement learning is to find a strategy $\pi$ or choose rules of action, to make the greatest value of the reward expected. However, in many practical issues people do not only demand the maximal reward, but also ensure that the price (cost) is not too more. It is various systems should operate more rationally and effectively under the conditions of security. The description of using reinforcement learning is constraint optimization problem of control within a certain condition and the greatest rewards [10] [11].

## II. RELATED WORK

There is credible work regarding self learning decision making for an agent through tails-and-error interaction. Ethan al. [1] helps us for comprehensive survey based on query based reinforcement learning in the form of agent learning reward and penalty without former in sequence about system. Most of the existing research work in reinforcement learning focus on improving the reward value and episodes in query based learning [3][12]. But in our knowledge agent should be fast searching in less amount time with respect to improving discount rate and number of episodes. Existing work on [6][10][13] agent self learning is based on three main approaches: agent on training, agent on work and neural network classifier used for removing redundant information in look-up-tale. This helps agent for exact learning. The difference between these previous works is not equal in the context of solutions tool like ANN [14][15], but more applicable in application- oriented approach for learning as well as searching. In these proposals, neural networks are used as decision helper or classifier. This paper contributes

towards design and simulating application-oriented RL searching algorithm for fast learning and capturing efficient

**Working procedure of RL**

Using this technique, the agent gets trained, who learn to behave intelligently and Agent's aim is to reach the goal. Now take the size of grid world is MxN, where M is equal to N as 10. The agent starts state position (1, 1) at the top left most. The goal is right most cells at the bottom are (10, 10). Agent can moves only one cell at a time to neighboring cell that is, up words, down words, to the right or to the left, unless the agent touches the border or wall, if the action is possible. When the agent is touches the border or wall, the action that makes the agent cross the border is not performed but it must remain stooped or take decision for next available action. This would be repeated until the agent reaches the goal.

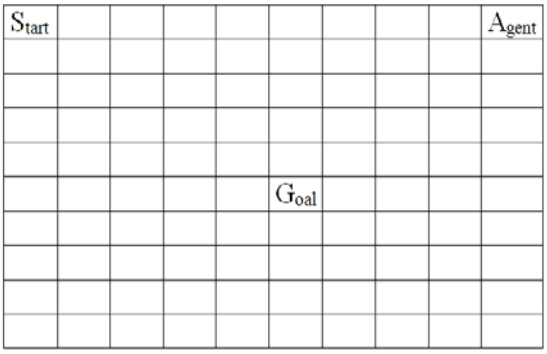



Fig.1.2: their size of the grid world 10x10 moved in the all direction left light up and down, assume if there movement at the pint (1, 1) and move towards down word (2, 1) and if the next action is to right, then moves to (1, 2). The cell, those agents are going to explore using a decision path is like a grid shown in fig.1.3. Agent finds decision path in fig.1.3,the start position to goal position in 10x10 grid world using Q-learning algorithm, agent select randomly one of four actions at a time.

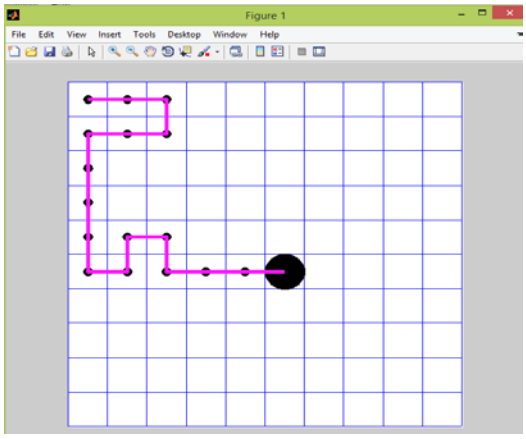



Fig.1.3: In the grid-world of 10x10, starting from (1, 1) agent move aiming the goal at (6, 6).

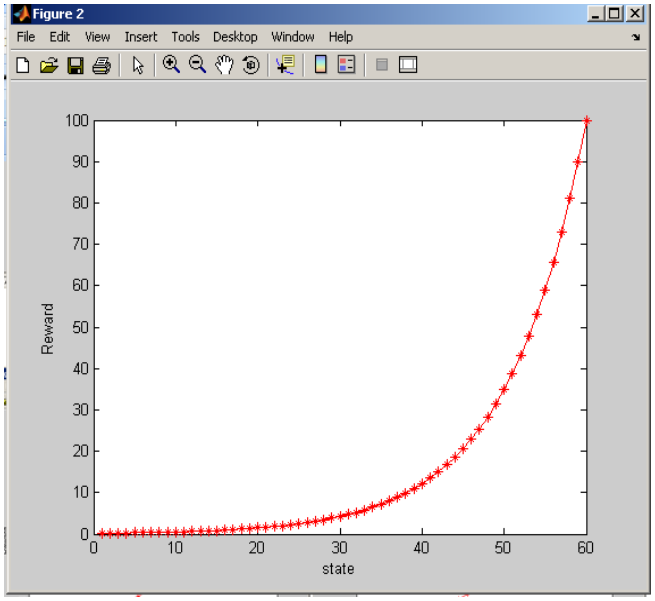
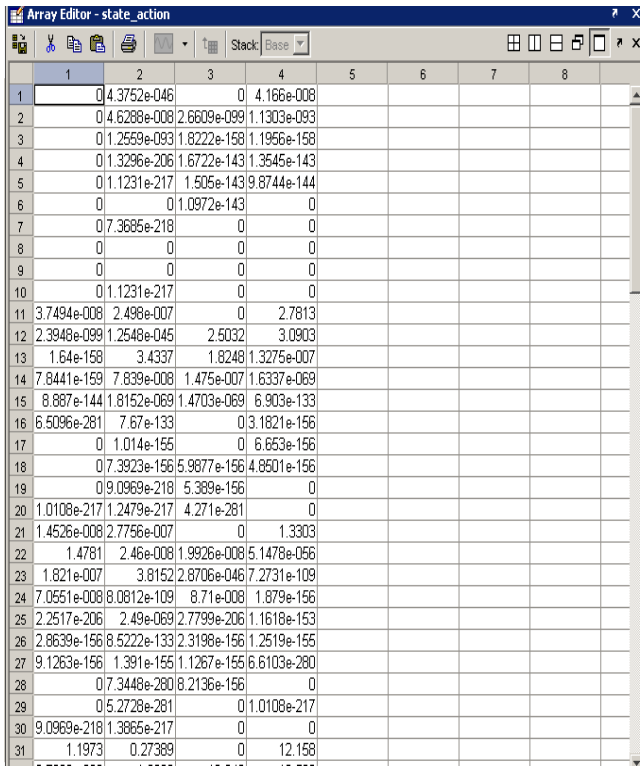



Fig.1.4: the corresponding performance graph

**Training Agent**

Train the agent [17] in the form of state-action pair. At this point you can view the learned agent in the form of Q-table. You can also check by clicking on run command line in the form of decision path. Now specify the inputs and goal state. State-action pair (100x4) implemented using 10x10 grid world. This shown below as follows

Table 4.1 State Action Selection with Q-Learning (Q-table)




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 4.3752e-046 | 0 | 4.166e-008 | | | | |
| 2 | 0 | 4.6288e-008 | 2.6609e-099 | 1.1303e-093 | | | | |
| 3 | 0 | 1.2559e-093 | 1.8222e-158 | 1.1956e-158 | | | | |
| 4 | 0 | 1.3296e-206 | 1.6722e-143 | 1.3545e-143 | | | | |
| 5 | 0 | 1.1231e-217 | 1.505e-143 | 9.8744e-144 | | | | |
| 6 | 0 | 0 | 1.0972e-143 | 0 | | | | |
| 7 | 0 | 7.3685e-218 | 0 | 0 | | | | |
| 8 | 0 | 0 | 0 | 0 | | | | |
| 9 | 0 | 0 | 0 | 0 | | | | |
| 10 | 0 | 1.1231e-217 | 0 | 0 | | | | |
| 11 | 3.7494e-008 | 2.498e-007 | 0 | 2.7813 | | | | |
| 12 | 2.3948e-099 | 1.2548e-045 | 2.5032 | 3.0903 | | | | |
| 13 | 1.64e-158 | 3.4337 | 1.8248 | 1.3275e-007 | | | | |
| 14 | 7.8441e-159 | 7.839e-008 | 1.475e-007 | 1.6337e-069 | | | | |
| 15 | 8.887e-144 | 1.8152e-069 | 1.4703e-069 | 6.903e-133 | | | | |
| 16 | 6.5096e-281 | 7.67e-133 | 0 | 3.1821e-156 | | | | |
| 17 | 0 | 1.014e-155 | 0 | 6.653e-156 | | | | |
| 18 | 0 | 7.3923e-156 | 5.9877e-156 | 4.8501e-156 | | | | |
| 19 | 0 | 9.0969e-218 | 5.389e-156 | 0 | | | | |
| 20 | 1.0108e-217 | 1.2479e-217 | 4.271e-281 | 0 | | | | |
| 21 | 1.4526e-008 | 2.7756e-007 | 0 | 1.3303 | | | | |
| 22 | 1.4781 | 2.46e-008 | 1.9926e-008 | 5.1478e-056 | | | | |
| 23 | 1.821e-007 | 3.8152 | 2.8706e-046 | 7.2731e-109 | | | | |
| 24 | 7.0551e-008 | 8.0812e-109 | 8.71e-008 | 1.879e-156 | | | | |
| 25 | 2.2517e-206 | 2.49e-069 | 2.7799e-206 | 1.1618e-153 | | | | |
| 26 | 2.8639e-156 | 8.5222e-133 | 2.3198e-156 | 1.2519e-155 | | | | |
| 27 | 9.1263e-156 | 1.391e-155 | 1.1267e-155 | 6.6103e-280 | | | | |
| 28 | 0 | 7.3448e-280 | 8.2136e-156 | 0 | | | | |
| 29 | 0 | 5.2728e-281 | 0 | 1.0108e-217 | | | | |
| 30 | 9.0969e-218 | 1.3865e-217 | 0 | 0 | | | | |
| 31 | 1.1973 | 0.27389 | 0 | 12.158 | | | | |

## III. CONCLUSION AND FUTURE WORK

This research paper, emphasis the survey and the working procedure of the RL enforcement technique how to make agent learn over grid world problem based on the reinforcement learning. This paper also illustrates that knowledge acquired by the agent from environment and corresponding performance between reward and state.

In the future work, it may improve using wireless sensor network and other decision techniques such as PCA and SVM in the form of decision classifiers.

## REFERENCES

[1] Ethan Alpayadin, “Introduction to machine learning”, MIT press Cambridge, 2005

[2] Jun Wang Carl Trooper,” Optimizing Time Warp Simulation with Reinforcement Learning Techniques”, IEEE, 2007, pp.577-584

[3] Lucian Robot and Bart,” A comprehensive survey of multi agent reinforcement learning”, vol-38, No.2, IEEE 2008, pp. 157-172

[4] Hitoshi Ima and Yaouk Karo, “Swarm Reinforcement Learning Algorithms Based on Sara Method”, IEEE, 2008, pp.2045-2049

[5] Habit Karbasian1, Maida N, “Improving Reinforcement Learning Using Temporal Deference Network EUROCON”, IEEE, 2009, pp.1716-1722

[6] Renate R., dam Silva, Claudio A,” An Enhancement of Relational Reinforcement Learning”, IEEE, 2008, pp.2055-2026

[7] Keita Halmahera, Tadahiro, “Effective integration of imitation learning and reinforcement learning by generating internal reward”, Eighth International Conference on Intelligent Systems Design and Applications, IEEE 2008pp.121-126

[8] Taraire Taniguchi,”Role differentiation process by division of reward function in multi agent reinforcement learning”, IEEE 2008, pp.387-393

[9] Yang and David Grace,” Cognitive Radio with Reinforcement Learning Applied to Heterogeneous Multicast Terrestrial Communication Systems”, IEEE, 2009, pp.1-6

[10] Wang Xian, Yang Chinua,” Decoupling control using a PSO-based Reinforcement Learning”, International Conference on Computational Intelligence and Natural Computing IEEE, 2009, pp.170-173

[11] Leslie, Pack Knelling,” Reinforcement Learning: A Survey”, Journal of artificial intelligence Research, 1996, pp. 237-277.

[12] Zhan Shang, Doan Hominy Chen,” Can Reinforcement Learning Always Provide the Best Policy”, IEEE, 2007, pp.224-228

[14] Zhao Jin, Wei Yi Liu, Jean, Jin,” Finding Shortcuts from Episode in Multi-agent Reinforcement Learning”, Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, IEEE, 2009, pp.2306-2311

[15] Tomohiro Yamaguchi, Takuma Nishimura, “How to recommend preferable solutions of user in interactive reinforcement learning”, The IEEE, 2008, pp.2050-2055

[16]. Tom Mitchell, “introduction to machine learning”, MIT presses Cambridge, 2005

[17] MATLAB 7.0.1 toolbox documentation help, especially network and GUI (Guide)