International Journal of Engineering & Technology, 7 (1.5) (2018) 274-278

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Implementation of modified SARSA learning technique in EMCAP

D. Ganesha [1], Vijayakumar Maragal Venkatamuni [2]

[1]Bharathiar University, Department of ISE, PVP Polytechnic,
Dr.AIT campus Outer Ring Road, Malathahalli, Nagarabhavi, Bangalore – 560056, Karnataka, India
[2]Department of Computer Science, Research Progress Review Committee[RPRC],
Dr. Ambedkar Institute of Technology, Visvesvaraya Technological University,
Bengaluru – 560056, Karnataka, India
**Corresponding author E-mail: ganesh207d@gmail.com*

Abstract

This research work presents analysis of Modified Sarsa learning algorithm. Modified Sarsa algorithm. State-Action-Reward-State-Action (SARSA) is an technique for learning a Markov decision process (MDP) strategy, used in for reinforcement learning int the field of artificial intelligence (AI) and machine learning (ML). The Modified SARSA Algorithm makes better actions to get better rewards. Experiment are conducted to evaluate the performace for each agent individually. For result comparison among different agent, the same statistics were collected. This work considered varied kind of agents in different level of architecture for experiment analysis. The Fungus world testbed has been considered for experiment which is has been implemented using SwI-Prolog 5.4.6. The fixed obstructs tend to be more versatile, to make a location that is specific to Fungus world testbed environment. The various parameters are introduced in an environment to test a agent's performance. This modified SARSA learning algorithm can be more suitable in EMCAP architecture. The experiments are conducted the modified SARSA Learning system gets more rewards compare to existing SARSA algorithm.

Keywords: *Self learning, Cognitive Control, sarsa Learning.*

1. Introduction

In reinforcement learning (RL) in et.al [1] a agent searches for a perfect control system for a back to back decision issue. Not at all like in managed learning in perspective of the way that numerous down to earth issues (case robot control, framework enhancement and amusement playing) dive in this gathering, creating efficient reinforcement learning techniques is basic to the advance of AI. At the point when the consecutive strategy issue is displayed as a MDP [2], the operator's arrangement is spoken to being a cumulative from all state it might presumably experience to a likelihood flow inside the accessible activities.

In few instances, the agent may utilize the environmental surroundings to its experience interacting to calculate a type of the MDP then calculate a policy optimal off-line preparation practices for powerful development [3]. When taking in a model simply is not achievable, the agent can in any case find an ideal strategy using temporal differnce techniques[4].

Every instance the agent responds, the reaction by utilized to upgrade quotes of its action parameter operation, which forecasts the future expected reward reduced will get if it requires confirmed action in a provided state. The behavior strategy, utilized to manage representative in learning process, is various from the computed strategy, whose parameter will be discovered under specific terms, Temporal difference methodology are assured in full converge into the restriction towards attaining the optimal function that is action parameter from which an optimal stategy can very quickly be expressed. In off tategy Temporal difference methodology techniques, for example Q-learning [5].The behavior strategy, utilized to manage the representative in learning process is various from the approximation strategy ,whose parameter will be discovered.

The benefit of the proposed method is representative container use a investigative conduct to ensure it gathers information which can be sufficiently diverse and on policy approach, where the estimation and behavior policies are identical, even offers advantages which can be essential. In specific, it offers more powerful convergence guarantees whenever coupled with work guess, since off-strategy methodologies can separate in that example [6] and has now an advantage of off-approach rehearses in its execution where as in online the estimation arrangement, that is iteratively upgraded, can be the approach which is utilized to control its conduct. By strengthening research over the long haul, on-strategy systems can reveal precisely the samewithin the limitation as off strategy techniques. The on stategy is a technique called classic Sarsa [7], which is ben known for its five elements used in its revision guideline: the present state & action s_t as well as, and instant reward r, plus the future state & action s_{t+1} and a_{t+1}. The application of a_{t+1} presents adjustment that is extra the change whenever approximation strategy is stochastic because is normally the truth for on policy practices like Sarsa.Although the algorithm that is ensuing which we call Modified Sarsa, may provide significant benefits over Sarsa, this has never ever been methodically examined and it is may be not trusted in training.

In this paper, we introduce a hypothetical and examination experimental of Sarsa. With respect to part hypothetical we demonstrate that Modified Sarsa stocks the union that is just like Sarsa and therefore discovers the perfect policy into the limitation under specific conditions. We additionally reveal that Modified Sarsa consumes reduced adjustment in its apprises than Sarsa and show which facts subscribe to this space. In general part the observed our model compares the outcome achived of Modified Sarsa utilizing the outcome of Sarsa what's more, Q-realizing. We figure twofold speculations as to the execution refinement among Modified Sarsa and both of these techniques and affirm them using two conditions that are standard the issue that is strolling the breezy matrix world issue.we outcomes which can be additionally current domain names which can be extra some great benefits of Modified Sarsa in a wider environment.

2. Multi-agent learning

An environment in multi - agent which there was multiple representative, where they connect to the other person and Further, where we can find limitations in that environment so that agents may well not at any given time understand everything in regards to the agents environmental test bed . A Multi agent learning address the problem domains, agents [8] involved is numerous. The search room considered is extraordinarily huge. A tiny alteration in learning actions may result in random often changes into the resultant macro-level properties of this team that is multi agent an entire as a result of interaction of the agents. Multi-agent learning involves learners being numerous each adapting and learning in the context of others [9].

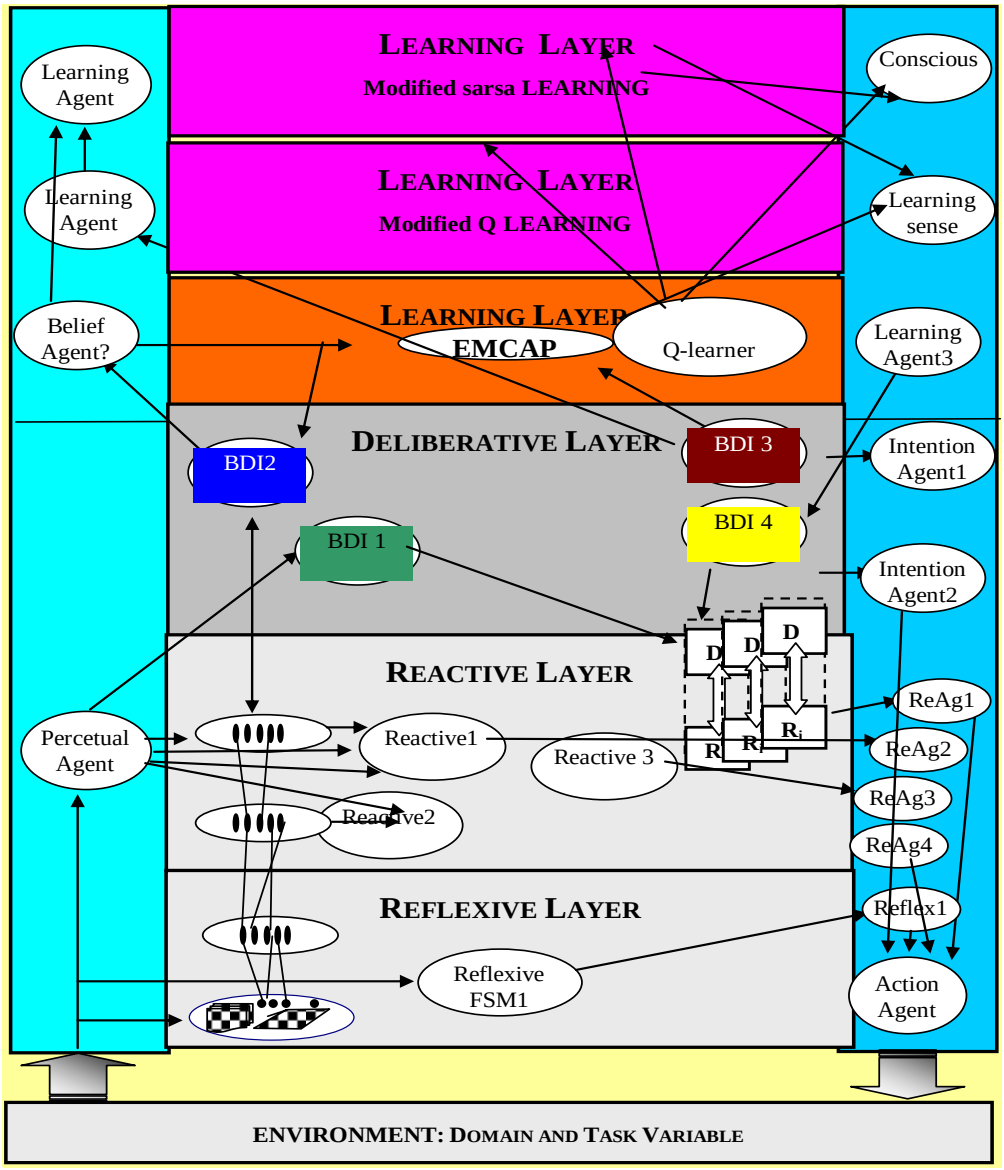

Fig. 1: Modified sarsa learning architecture

Parallelism, scalability, easier construction and price effectiveness are primary faculties of multi-agent systems. Having these characteristics, multi-agent systems are accustomed to resolving complex dilemmas, search in big domain names, perform advanced tasks, and work out more fault-tolerant and systems [9] that is dependable. Generally in most of these functional systems being existing agents' behavior and coordination schemes are made and fixed by the designer. But a real agent with incomplete and fixed behavior and knowledge may not be acceptably efficient in a powerful, complex or an environment that is changing. Consequently to own all advantages of using something that is an agent that is multi-agent must learn how to handle the new, hidden and dynamic circumstances. In roughly all the groups being current are multi-agent, agents learns independently [9].

3. Background

The decision that is sequential spoke in reinforcement learning are often expressed as Markov decision process, which can be expressed as 4-rows $\langle S, A, T, R \rangle$ where
S is the arrangement of every proceeding with express the operator can experience,
An is the arrangement of all activities accessible,

$$T\left(s, a, s^{'} = p(s'/s, a')\right)$$

$$R\left(s, a, s^{'} = E(r/s, a, s')\right)$$

R is the reward operation, and T is the transition operation.
The objective of agent is to establish/identify a strategy that is $\pi^* = p(a/s)$ optimal that maximizes the modified reward return:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k r_t + k + 1$$

Where γ is a reward influence with $0 \le \gamma \le 1$ All temporal difference technique are based on computed parameter $V^{\Pi} = (s)$ operations. The operation that is state-value the modified return when the agent is in the states and follows the strategy. The action value operation π gives the modified return when the agent takes action a in a strategy s and follows stategy post completion. These two operations are connected by

$$V^{\Pi} = (s) = \sum \pi(s,a) Q^{\pi}(s,a)$$

Temporal difference technique search for the optimal function that is action parameter $Q^* = (s_t, a_t)$ from which an optimal strategy can simply π be inferred. $Q^* = (s_t, a_t)$ Can be established by continously informing the compute $Q(s_t,a_t)$. The method that strategy updates its Q values using the update rule

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_{t+1} + \gamma \max_a Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\right]$$

The administrator strategy to be covetous, which ensures the Q esteems focalize to $Q(s_t,a_t)$. The conduct approach of Q-learning is normally construct and exploratory with respect to $Q(s_t,a_t)$. The apprise condition of Sarsa is For Sarsa the conduct arrangement and the computation strategy are equivalent. The apprise condition of Sarsa is

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\right]$$

Since Sarsa is on-approach, it won't meet to q that is ideal as long as investigation happens. In any case, by tempering investigation after some time, Sarsa can join to q that is ideal, much the same as Q-learning.

4. Modified sarsa

Since Sarsa merging assurance requires that each reliable state be looked at unendingly, the conduct and moreover the estimation strategy is typically stochastic to have the capacity to ensure research that is enough. Being a total outcome, there is a variance that is significant Sarsa updates, since a_{t+1} isn't chosen deterministically. But, the variance that is extra by Sarsa comes from the insurance strategy stochasticity, which will be recognized to representative[10]. Modified Sarsa is just a variety of Sarsa that misuses learning to stop stochastic into the strategy from additional variation that is increasing. Therefore by improvement, $Q(s_{t+1}, a_{t+1})$ but on its parameter that is modified E.
The resulting update rule is:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_{t+1} + \gamma \sum_a \pi(S_{t+1}, a) Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\right]$$

Algorithm 1, demonstrations the Modified Sarsa technique is complete. Since the revision guideline of Modified Sarsa will not utilize action consumed s_{t+1}, action selection may appear following the up-date. Doing this is beneficial in dilemmas states which are containing actions which are coming back in other words. $P(s_{t+1} = s_t) > 0$. When $s_{t+1} = s_t$, doing an change of $Q(s_t,at)$, may also appraise $Q(s_{t+1}, a_t)$, attaining an improved compute before action choice happens.

Algorithm 1: Modified Sarsa

1: Initialize Q (s, a) arbitrarily for all s, a
2: loop {over episodes}
3: Initialize s
4: repeat {for each step in the episode}
5: choose a from s using policy π derived from Q
6: take action a, observer and s'
7. $V_{s'} = \sum a\pi(s', a) Q(s', a)$
8 $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma V_{s'} - Q(s,a)]$
9: $s \leftarrow s'$
10: until s is terminal
11: end loop
If the policy π is greedy, π (s,a)for all a except for the action for which Q has its maximal value. Therefore, in the case of a greedy policy (6) simplifies to
$V(s) = \max Q(s,a)$

Therefore, Q-learning's change guideline (3) is a complete instance that is unique of Sarsa's enhances guideline (5) for the way it is once the computes policy is greedy. Nevertheless, the Modified Sarsa is completely compared to Q-learning ,.since the previous is on strategy and also the future case is off strategy.

5. Experimental setup

The proposed sarsa algorithm ended up being tested in a term using fungus by utilizing SWI-Prolog.the fungus happens to be meant to have both energy and fixed (Figure 1). The obstructs which can be fixed a lot more versatile, to generate a location that is certain of environment. They will have various parameters which are different the surroundings for the agent's locomotive for checking the performance .The experimental environment is made through the checkbox ,standard fungus ,little fungus, bad fungus ,golden ore, crystal ,medicine. The agents are manufactured into the environment by using Prolog . All the parameters are changed according to needs that can be experimentally because they are defined in a setup module. The agent cannot distinguish among standard fungus, small fungus, and bad fungus until it gathers or uses them. The trials were performed for the tantamount number of specialists, and the sort of the indistinguishable practically identical amount of growths (counting standard, little, and terrible), mineral (counting standard and brilliant metal) with similar things

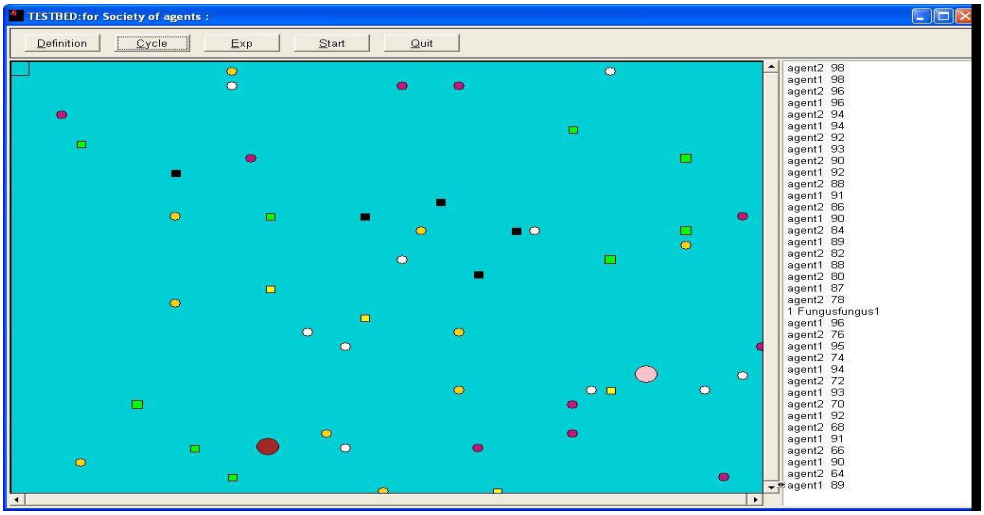

Fig. 2: Fungus world Testbed

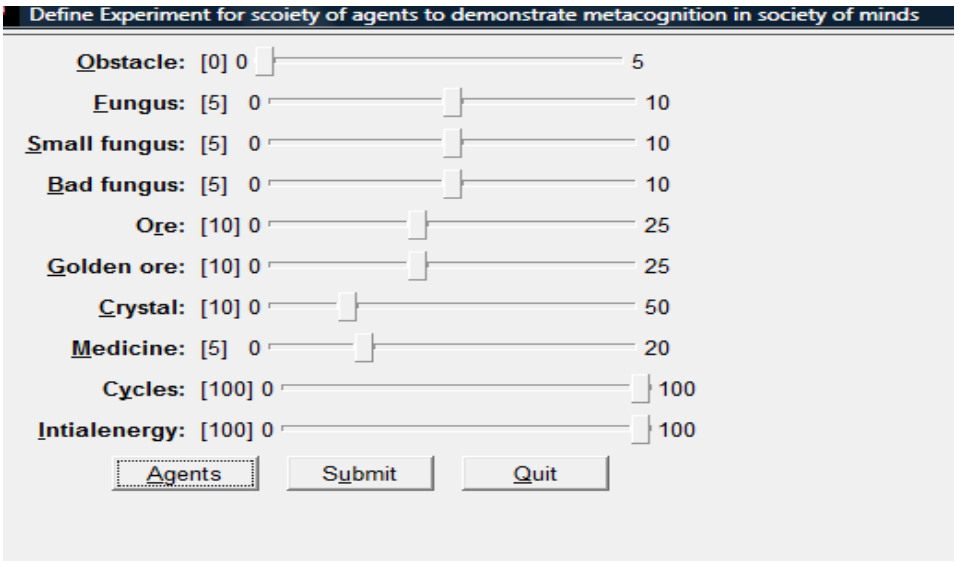

Fig. 3: Types of foods

Standard Fungus: Fungus is a supplement for the the agents. Every fungus that is standard a property that is bona fide 10 vitality items. At first, every specialist has foreordained vitality products. The agent runs on the difficulty and the quantity that is quick of products for each and every duration that is solitary. Fungus: the fungus that is little a realtor 5 energy products. In case agent makes use of a fungus that is small 5 energy products (standard) are placed into the vigor space for storing . Bad Fungus: the fungus that is bad 0 energy products. In case representative uses fungus that is bad it gets energy that is null. Furthermore, depraved fungus upsurges the absorption cost, changes the k calorie burning influence. Ore: The gathering of ore may be the goal that is fundamentally of the agent. Both agent team attempts toward gather as abundant mineral by way of likely into the environment. Golden Ore variety of golden ore upsurges the presentation of the agent. One bit that is smaller of ore is corresponding to five ore that is Crystal that is standard of crystal advances the presentation of agent through having a component that remains dull when compared with ore. Drugs: the medication impacts the metabolic rate associated with representative into the testbed. The assortment of medication decreases the metabolic rate.

Table for learning

Time(t)	SMCA performance	EMCAP with sarasa learning
0	0	0
1	1	2
2	4	7
3	9	11
4	14	19
5	20	21
6	21	26
7	27	31
8	30	39
9	34	43
10	40	49

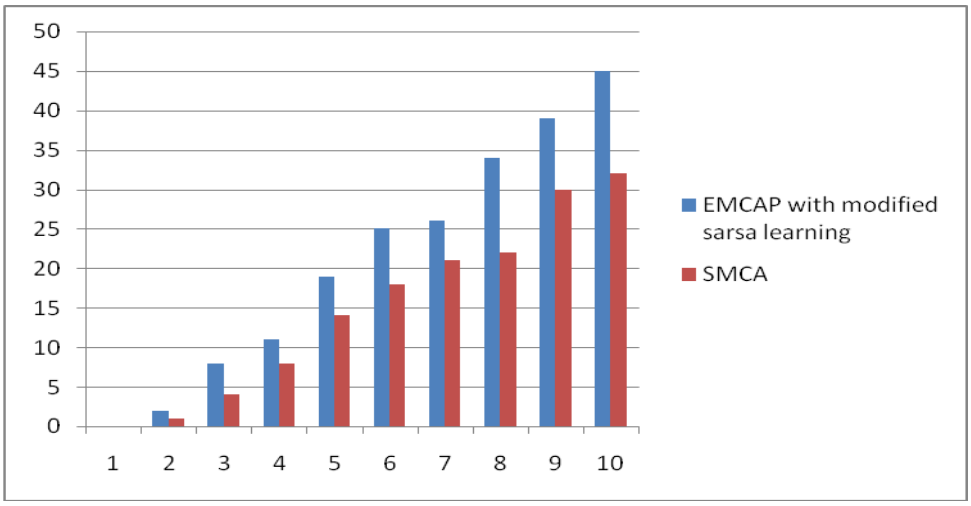

Fig. 4: Graphs

.

The 6th layer in EMCAP sarsa Learning changes accommodation authoritative at one amount about accomplishments at addition amount for tasks. The modified learning determines considering exactly what and exactly how to map situations to action for making the most of a reward. Modified system that is the sarsa Learning or attempts to locate a reward that is optimum in the use of fungus for the action. The stimulus is described by the policy- response rules for representative behavior. Policy is really a core element of reinforcement learning. If the agent's action is a reward that is lower than policy will likely to be changed to another, also is actively seeking the reward that is high. Benefits determine the desirability and instant that is intrinsically of states. The very best actions for the representative are discovered by the test and mistakes. The representative is relocated to do the duty, utilizing the modified sarsa algorithm that is performed that is learning calculated with respect to the reward. the reward is determined with regards to the levels of energy regarding the representation. The performance associated with the modified sarsa algorithm that is learning about EMCAP architecture is calculated with SMCA that has shown in the graph.

6. Conclusion

This research adopted the modified SARSA learning algorithm for EMCAP. This exploration paper showed how to outline and receive an modified SARSA learning system in subjective models in the wide region of Artificial Intelligence. This paper additionally offered how to outline rulesfor modified SARSA algorithm. This paper investigates about the modified SARSA-learning algorithm for EMCAP.

Experiments are conducted on agents in EMCAP using Fungus world testbed [1]. Agents from different levels of Architecture were employed for this experiment. In this experiment to show the effect of modified SARSA-learning on an agent, we considered comparisons between SRASA and modified SARSA-learning and the agents are experimented with high percentage life expectancy and resources (refer graph).The proposed algorithm in a SARSA has been simulated by using a prolog. The modified SARSA-learning technique had been tested in a Fungus world environment. The modified SARSA learning algorithm gets more rewards and gives more performance as shown in the graph.

References

[1] L. P. Kaelbling, M. L. Littman, and A. P. Moore, “Reinforcementlearning: A survey,”Journal of Artificial Intelligence Research, vol. 4,pp. 237–285, 1996.

[2] R. E. Bellman, “A Markov decision process,” Journal of MathematicalMechanics, vol. 6, pp. 679–684, 1957.

[3] R. E. Bellman, Dynamic Programming. Princeton, NJ.: PrincetonUniversity Press, 1957.

[4] R. S. Sutton, “Learning to predict by the methods of temporaldifferences,” Machine Learning, vol. 3, pp. 9–44, 1988.

[5] C. Watkins and P. Dayan, “Q-learning,” Machine Learning, vol. 8 (3-4),pp. 9–44, 1992.

[6] J. A. Boyan and A. W. Moore, “Generalization in reinforcement learning: Safely approximating the value function,” in Advances inNeural Information Processing Systems, pp. 369–376, 1995.

[7] G. Rummery and M. Niranjan, “On-line Q-learning using connectionistsystems,” Cambridge University, Tech. Rep. CUED/F-INFENG/TR166, 1994.

[8] Dr. Hamid R. Berenji David Vengerov “Learning, Cooperation, andCoordination in Multi-Agent Systems”, in Proceedings of 9th IEEE Int.Conf.On Fuzzy Systems.

[9] Deepak A. Vidhate, ParagKulkarni “Enhancement in Decision Makingwith Improved Performance by Multiagent Learning Algorithms”IOSR Journal of Computer Engineering, Vol.1, issue 18, pp18-25,2016.

[10] L. Tokarchuk 1 , J. Bigham 1 and L. Cuthbert “Fuzzy Sarsa: An approach to fuzzifying Sarsa Learning 2015.

[11] T. Padmapriya and V. Saminadan, “Inter-cell Load Balancing technique for multi-class traffic in MIMO-LTE-A Networks”, International Journal of Electrical, Electronics and Data Communication (IJEEDC), ISSN: 2320- 2084, vol.3, no.8, pp. 22-26, Aug 2015.

[12] S.V.Manikanthan and K.Baskaran “Low Cost VLSI Design Implementation of Sorting Network for ACSFD in Wireless Sensor Network”, CiiT International Journal of Programmable Device Circuits and Systems,Print: ISSN 0974 – 973X & Online: ISSN 0974 – 9624, Issue : November 2011, PDCS112011008.

[13] Rajesh, M., and J. M. Gnanasekar. "Congestion control in heterogeneous wireless ad hoc network using FRCC." Australian Journal of Basic and Applied Sciences 9.7 (2015): 698-702.