Multi-agent Cooperation Models by Reinforcement Learning (MCMRL)

Deepak A. Vidhate
Research Scholar, Department of Computer Engineering, College of Engineering, Shivajinagar, Pune, Maharashtra, India

Parag Kulkarni, PhD
CEO & Chief Scientist,
iKnowlation Research Lab. Pvt. Ltd., Shivajinagar, Pune, Maharashtra, India

ABSTRACT
A novel approach to multi-agent cooperation methods by reinforcement learning (MCMRL) is proposed in this paper. Cooperation methods for reinforcement learning depend on the multi-agent scheme are proposed and implemented. Different cooperation methods of cooperative reinforcement learning of each agent proposed here i.e. group method, dynamic method, goal-oriented method. Implementation results have demonstrated that the suggested cooperation methods are capable to accelerate the aggregation of agents that accomplish best action strategies. This approach is developed for dynamic product availability in a three retailer shop in the market. Retailers can cooperate with each other and can get the benefit of cooperative information from their own policies that accurately represent their goals and interests. The retailers are the learning agents in the problem and apply reinforcement learning to learn cooperatively in the situation. By making the considerable theory of the dealer's inventory strategy, refill period, and entry procedure of the customers, the problem turns out to be Markov decision process model thus facilitating to apply learning algorithms.

General Terms
Computer Science→Artificial Intelligence→Machine Learning→Reinforcement Learning

Keywords
Cooperation methods, Dynamic buyer behavior, Multi-agent learning, Reinforcement learning

1. INTRODUCTION
Hundreds of shops across a region retailing thousands of products to millions of buyers are a good model of market chains. The sale point of each retailer confirms the information on each transfer i.e. date, buyer ID code, products purchased and their spent sum. This naturally yields a huge amount of records every day. If accumulated records are analyzed and turned into information then it becomes useful so that we can utilize an illustration to build forecasts. We can only gather information and expect to take out the answers to questions on data [1]. It is considered that it presents a procedure to facilitate the demonstration the observed data. Even if it is not known the highlights of the procedure responsible for the creation of records – for instance, buyer behavior – it is known that it is not totally accidental. The public does not walk to markets and purchase items at casual. We may be unable to recognize the procedure totally, but still, we can build a *useful and good approximation*. These temporary computations might not give details of everything, but may still be able to construct for some part of the data. There are many real-world problems that engage more than one thing for maximization of results [2].Retailers have always encountered the difficulty of sale the right goods that would produce the highest income for them. Finding the right products for a buyer or a service is a difficult task. In forthcoming time, retailers would suggest special package, simply customized for each purchaser, simply for the instant on the whole thing (correct item to the correct purchaser at the correct period) [3]. Different parameters need to be considered in this: variation in seasons, the dependency of items, special schemes, discount, and market conditions. Retailers can cooperate with each other for yield maximization in different situations [4]. A market model in the perspective of dynamic buyer behavior is studied in this paper. The following are the exact value addition of this paper.

Three seller retail stores are considered which sell a selected product and gives quantity concessions for customers purchasing many items. Seller's inventory strategy, refill period, and the entry procedure of the customers are measured. A Markov Decision Process (MDP) model is suggested for this system. A new way for context-based dynamic decision making by cooperative multi-agent learning algorithms is proposed. A novel move toward multi-agent cooperation methods by reinforcement learning (MCMRL) is proposed here. Communication methods for reinforcement learning build on the multi-agent scheme is proposed and implemented [5],[6]. The paper is ordered as given. Section 2, describes an innovative approach towards multi-agent cooperation methods by reinforcement learning (MCMRL). Section 3, illustrates the system kinetics of retail shops modeled by Markov decision procedure. Section 4 describes a simulation results all four methods with continuing price as the profit parameter. Section 5 describes concluding remark.

2. MULTI-AGENT COOPERATION METHODS BY REINFORCEMENT LEARNING (MCMRL)
The communication in multi-agent reinforcement can construct an advanced set of performances gained from the agents' proceedings. A piece of performance group (i.e. a universal action plan) is distributed among the agents via a *Limited Action Plan* (Q_i) [6]. Typically such limited strategies hold partial knowledge about the situation. These policies can be integrated to enhance the total of the incomplete reinforcements achieved using adequate communication model. The action strategies are produced by means of multi-agent Q-learning algorithm by collecting the reinforcements and building the agents to move towards the best plan Q*. When strategies $Q_1,\ldots,Q_x$ are integrated, it is feasible to develop new plan that is *Universal Action Plan* (*UAP* = $\{UAP_1,\ldots,UAP_x\}$), in which UAP_i indicates the **excellent rewards** obtained by agent i throughout the learning method [7]. Algorithm 1 describes a *share_plan* algorithm that distributes the agents' learning details. The plans are calculated by the Q-learning algorithm for each model. Excellent rewards are given toward *UAP* that forms a collection of the excellent accumulated reinforcements by the agents. These reinforcements will be again distributed by

means of the additional agents [7], [8]. Cooperation is carried out by the transformation of limited reinforcements as *UAP* is predicted by means of the excellent rewards. A *value* function is used to find out the best policy among the early states and target state for a given strategy that estimates *UAP* with the best rewards. The value function is determined by counting of stages the agent required to arrive at the target-state and the total of the obtained values in the policy among every start state and the target state [9].

Algorithm 1: Multi-agent Reinforcement Learning Algorithm

Algorithm *share_plan* (I, technique)
1. Initialization $Q_i(s, a)$ and $UAP_i(s, a)$
2. Communication by means of the agents $i \in I$;
3. Agents cooperate till the target state is found;
episode ← episode +1
4. Renew rule which calculates the reinforcement value;
$Q(s, a) \leftarrow Q(s, a) + \alpha (r + \gamma Q(s', a') - Q(s, a))$
5. Fcooperate (episode, technique, s, a, i);
6. $Q_i \leftarrow$UAP that is Q_i of agent $i \in I$ is modified by means of UAP_i.

The cooperate task choose a cooperation method. the episode, technique, s, a, I are the parameters, in which episode is an existing iteration, coordination technique is {group, dynamic, goal-oriented}, s and a are state and action selected correspondingly [9], [10];

2.1 Cooperation Methods

Different cooperation methods for cooperative reinforcement learning are proposed here:

i) *Group method* – rewards are distributed in a sequence of steps.
ii) *Dynamic method* – rewards are distributed in each action.
iii) *Goal-oriented method* – distributing the sum of rewards when the agent reaches the goal-state (S_{goal}).

Algorithm 2 Cooperation model

*Fcooperate (episode, technique,s,a,i) /*cooperation between agents as four cases*/*

q : count of sequence
1. Switch between cases
2. In case of Group method
if episode mod q = 0 then
get_Policy(Q_i, Q,UAP_i);*
3. In case of Dynamic method
$r \leftarrow \sum_{j=1}^{x} Qj(s, a)$;
$Q_i(s,a) \leftarrow r$;
get_Policy(Q_i, Q,UAP_i);*
4. In case of Goal-oriented method
if S = S_{goal} then
$r \leftarrow \sum_{j=1}^{x} Qj(s, a)$;
$Q_i(s,a) \leftarrow r$;
get_Policy(Q_i,Q,UAP_i);*

Algorithm 3 get_Policy
Function get_Policy(Q_i, Q,UAP_i) /* find out universal agent policy */*
1. for loop for each agent $i \in I$
2. for loop for each state $s \in S$
3. if value(Q_i, s) $\leq$ value(Q,s) then*
$UAP_i(s,a) \leftarrow Q_i(s,a)$;
4. end for loop

Group Method: Agents gather Expertness based reinforcement acquired from its actions during the learning progression. At the last part of the sequence (step q), every agent throws out the cost of Q_j to *UAP*. If reinforcement cost is appropriate, that is it enhances the effectiveness of another agent for given state the agents will afterward contribute to these expertise base reinforcements. The agent will persist to utilize its reinforcement with the intention for congregating latest values [11], [12].

Dynamic method: The communication in the *dynamic method* is obtained as: every action performed by agent generates a reward cost (+ or -), that is total of collected expertness based reinforcements to all agents to action *an* achieved in state *s.* Every agent cooperates to maximize the reinforcement value complying its own policy [12], [13].

Goal-Oriented method: The coordination happens as agent arrives at its target-state. Agent cooperates through situation intended to congregate a maximum number of expertness based rewards. It is essential for the reason that in the *Goal-oriented method* the agent distributes its reinforcement with a changeable count of occurrences. This coordination method utilizes as a quick collection of reinforcements collected by the agent during the communication. As soon as agent arrives at a target-state it throws out the cost of obtained reinforcements in a situation to the *UAP* [14].

3. MODEL DESIGN

The case with wedding season is considered for the development. Beginning from deciding the venue, booking the caterers, decoration, invitation cards, photography, beautician, cosmetics, household items, gifts, shopping of clothes, jewelry and other accessories for bride and groom, so many activities are involved [14] [15]. Such seasonable situations can be realistically implemented as follows: Customer who would go for clothing shop certainly will buy jewelry, footwear, and other accessories. Retailers of different products can come together and jointly satisfy customer requirements and would achieve the benefit of an increase in the product sale [15] [16]. Figure 1 gives a diagrammatic representation of the system.

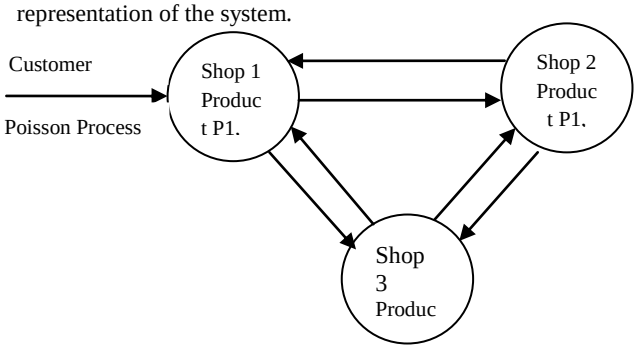

Fig. 1 A retail store model with three retailers

Below are mathematical notations for above model.

Consumers enter at the market by following a Poisson process with rate λ.

The seller posts per unit product price p to the incoming customers.

The seller has limited stock capacity I_{max} and follows a fixed reorder policy for refilling;

States:
Assume maximum stock level at each shop= I_{max}= i1,i2,i3= 20

State for agent 1 become (x_1,i_1) e.g. (5,0) that means 5 customer requests with 0 stock in shop 1. State for agent 2 become (x_2,i_2), State for agent 3 become (x_3,i_3),

State of the system become **Input** as (x_i,i_i)

Actions:
Assume set of possible actions i.e. action set for agent 1 is (that means Price of products in shop 1), A1 = Price p = {8 to 14} = {8.0;9.0; 10.0; 10.5; 11.0; 11.5; 12.0; 12.5; 13.0; 13.5}. Set of possible actions i.e. action set for agent 2 is A2 = Price p = {5 to 9} = {5.0;6.0; 7.0; 7.5; 8.0; 8.5; 9.0}. Set of possible actions i.e. action set for agent 3 is A3 = Price p = {10 to 13} = {10.0; 10.5; 11.0; 11.5; 12.0; 12.5; 13.0}.

The output is the possible action taken i.e. price in this case. It is now the state-action pair system can be easily modeled using Q learning i.e. **Q(s, a).**

Whenever a customer places a request for a product, a decision needs to be made regarding whether to accept or deny the request. Another retailer observes the action taken by the first retailer and be prepared to sell his product. In this way as a sale in one shop increases automatically other shops get informed so they can sell their products.

4. RESULTS

Algorithms are tested on one year's transaction dataset of three different retail shops and results are observed.

The ***group method*** appears to be extremely strong converging very fast to an optimal action policy 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 *group method,* the global policy converges to the best action strategy as there is an intermission of series necessary to gather good reinforcements.

The global action strategy of the ***dynamic method*** is able to gather good reward values in small learning series. It is observed that after some series, the performance of global strategy reduces. This takes place because the states neighboring to the goal state begin to gather much higher reward values giving to a local maximum. It punishes the agent because it will no longer stay in the other states. In the ***dynamic method***, as the RL algorithm renews learning values, actions with higher gathered rewards are chosen with top probability than actions with low gathered reinforcements.

In the ***goal-driven method***, the agent distributes its learning in a changeable number of sequences and the cooperation acquired when the agent arrives at the goal-state. The global action strategy of the goal-driven method is able to gather good reward values, given that there is a sum of iteration series to gather values of acceptable rewards. The performance of the cooperative learning algorithms is generally small in the early series of the learning process with the goal-driven model.

Figure 2,3,4 respectively shows that profit margin vs a number of states given by simple Q learning (without cooperation) and group, dynamic and goal oriented methods (with cooperation). Profit obtained by the cooperative methods i.e. group, dynamic and goal oriented methods is much more than that of without cooperation method i.e. simple Q learning for agent 1 in the multi-agent scenario.

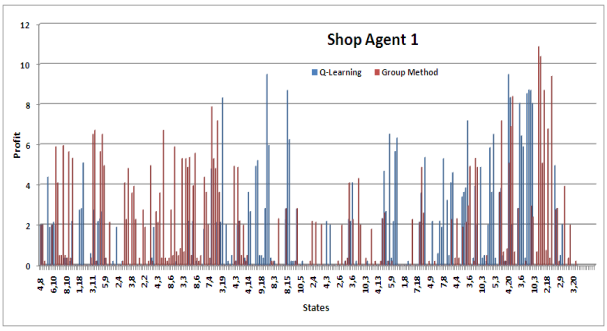

Fig. 2 States vs Profit for Agent 1 by Q-learning & Group Method

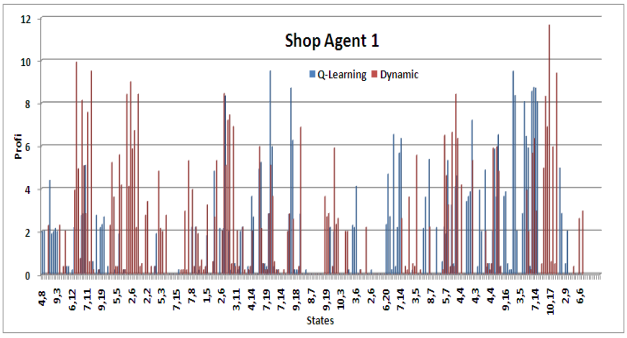

Fig. 3 States vs Profit for Agent 1 by Q-learning & Dynamic Method

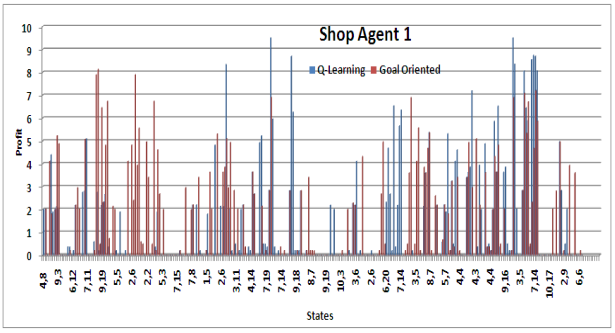

Fig. 4 States vs Profit for Agent 1 by Q-learning & Goal Oriented Method

Figure 5,6,7 respectively shows that profit margin vs a number of states given by simple Q learning (without cooperation) and group, dynamic and goal oriented methods (with cooperation). Profit obtained by the cooperative methods i.e. group, dynamic and goal oriented methods is much more than that of without cooperation method i.e. simple Q learning for agent 2 in the multi-agent scenario.

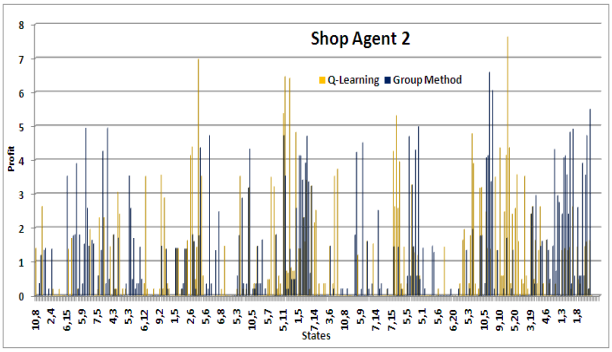

Fig. 5 States vs Profit for Agent 2 by Q-learning & Group Method

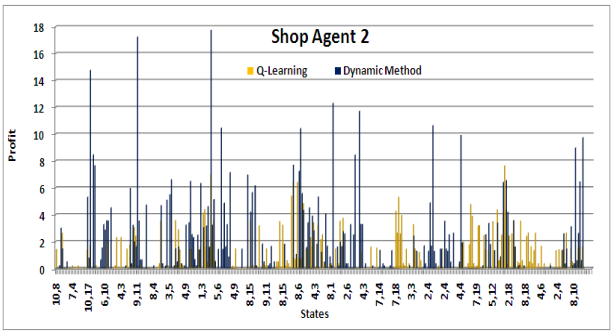

Fig. 6 States vs Profit for Agent 2 by Q-learning & Dynamic Method

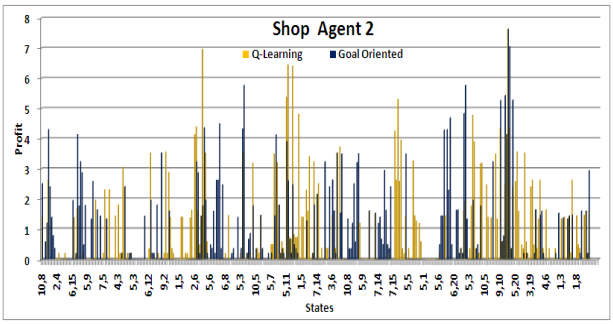

Fig. 7 States vs Profit for Agent 2 by Q-learning & Goal Oriented Method

Figure 8,9,10 respectively shows that profit margin vs a number of states given by simple Q learning (without cooperation) and group, dynamic and goal oriented methods (with cooperation). Profit obtained by the cooperative methods i.e. group, dynamic and goal oriented methods is much more than that of without cooperation method i.e. simple Q learning for agent 2 in the multi-agent scenario.

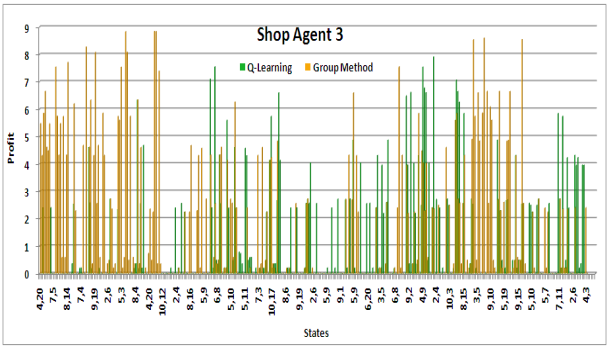

Fig. 8 States vs Profit for Agent 3 by Q-learning & Group Method

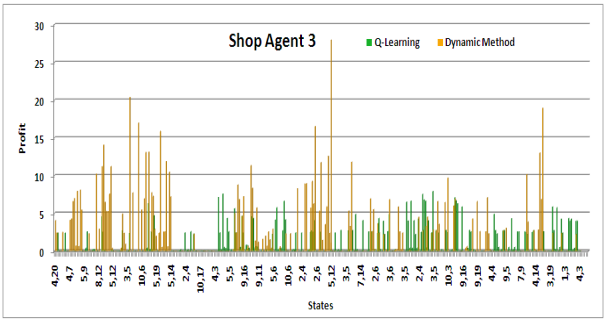

Fig. 9 States vs Profit for Agent 3 by Q-learning & Dynamic Method

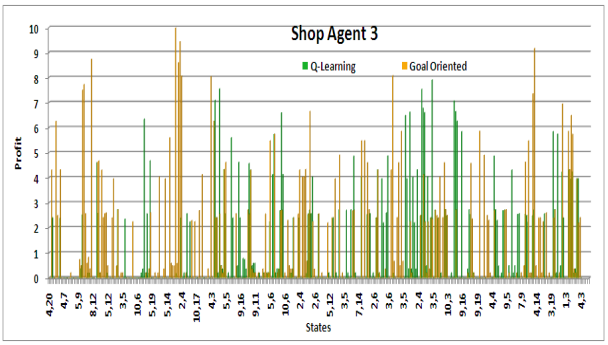

Fig. 10 States vs Profit for Agent 3 by Q-learning & Goal Oriented Method

Figure 11 shows the graphical analysis of the results obtained by four methods in four different quarters in one year. Figure 11 is described as: agent 1 gets maximum profit in 4th quarter using Q learning, group, and dynamic methods and it gets maximum profit in 2nd quarter using the goal-oriented method. Agent 2 gets maximum profit in 1st,2nd and 3rd quarter using dynamic method whereas it gets average profit in 4th quarter using Q learning, group and goal oriented method. Agent 3 gets maximum profit in 1st, 3rd and 4th quarter using dynamic method whereas it gets average profit in 2nd quarter using Q learning, group and goal oriented method.

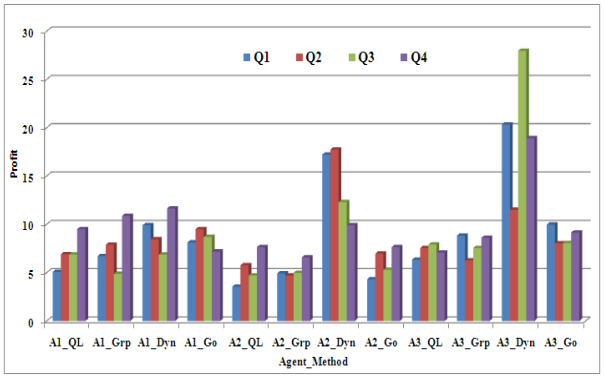

Fig. 11 Quarterly Profit obtained by all shop agents by four learning methods

5. CONCLUSION

Dynamic customer behavior is clearly understood using the new approach. The results obtained by the projected cooperation methods show that such methods can put into a quick convergence of agents that interchange the rewards. It also shows that cooperative methods gives a healthy performance in high-density, incompletely and composite situation. It provides a help to the interchange of best rewards to obtain a good universal action plan. All cooperation methods are able to guarantee best rewards which were acquired along learning process and change with a group of best rewards received in incomplete action strategies. By knowing the exchanging the Q function through four different methods i.e. group, dynamic, goal-oriented and expert agent method, the shop agent calculate best probable product that gives maximum profit to it. Multi-agent cooperation methods by reinforcement learning (MCMRL) shows that such methods can put into a fast linking of agents that replace rewards.

6. REFERENCES

[1] Deepak A. Vidhate and Parag Kulkarni, "Expertise Based Cooperative Reinforcement Learning Methods (ECRLM)", *International Conference on Information & Communication Technology for Intelligent System, Springer book series Smart Innovation, Systems and Technologies* (SIST, volume 84), Cham, pp 350-360, 2017

[2] L. Raju Chinthalapati, Narahari Yadati, And Ravikumar Karumanchi, "Learning Dynamic Prices In Multi-Seller Electronic Retail Markets With Price Sensitive Customers, Stochastic Demands, And Inventory Replenishments", *IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews*, Vol. 36, No. 1, January 2008

[3] Deepak A. Vidhate and Parag Kulkarni, "New Approach for Advanced Cooperative Learning Algorithms using RL methods (ACLA)" *VisionNet'16 Proceedings of the Third International Symposium on Computer Vision and the Internet*, ACM DL pp 12-20, 2016.

[4] Young-Cheol Choi, Student Member, Hyo-Sung Ahn "A Survey on Multi-Agent Reinforcement Learning: Coordination Problems", *IEEE/ASME International Conference on Mechatronics and Embedded Systems and Applications*, pp. 81 – 86, 2010.

[5] Deepak A. Vidhate and Parag Kulkarni, "Innovative Approach Towards Cooperation Models for Multi-agent Reinforcement Learning (CMMARL) "*International Conference on Smart Trends for Information Technology and Computer Communications* Springer, Singapore, 2016 pp. 468-478.

[6] Zahra Abbasi, Mohammad Ali Abbasi "Reinforcement Distribution in a Team of Cooperative Q-learning Agent", *Proceedings of the 9th ACIS Int. Con. on Software Engineering, Artificial Intelligence, and Parallel/Distributed Computing* 978-0-7695-3263-9/08 pp 154-160, IEEE 2008.

[7] Deepak A. Vidhate and Parag Kulkarni, "Implementation of Multi-agent Learning Algorithms for Improved Decision Making", *International Journal of Computer Trends and Technology (IJCTT)*, Volume 35 Number 2-May 2016

[8] Li-mei GAO, Jun ZENG, Jie WU, Min LI "Cooperative Reinforcement Learning Algorithm to Distributed Power System based on Multi-Agent" *3rd International Conference on Power Electronics Systems and Applications* Digital Reference: K210509035, 2009

[9] Deepak A. Vidhate and Parag Kulkarni, "Enhancement in Decision Making with Improved Performance by Multi-agent Learning Algorithms" *IOSR Journal of Computer Engineering*, Volume 1, Issue 18, pp 18-25, 2016.

[10] Adnan M. Al-Khatib "Cooperative Machine Learning Method" *World of Computer Science and Information Technology Journal (WCSIT)* ISSN:2221-0741 Vol.1, 380-383, 2011.

[11] Deepak A. Vidhate and Parag Kulkarni, "Multilevel Relationship Algorithm for Association Rule Mining used for Cooperative Learning*", International Journal of Computer Applications* (0975 – 8887), volume 86, number 4, pp 20--27,2014

[12] Liviu Panait, Sean Luke "Cooperative Multi-Agent Learning: The State of the Art", *Journal of Autonomous Agents and Multi-Agent Systems*, 11, 387–434, 2005.

[13] Jun-Yuan Tao, De-Sheng Li "Cooperative Strategy Learning In Multi-Agent Environment With Continuous State Space", *IEEE Int. Conf. on Machine Learning*, 2006.

[14] Deepak A. Vidhate and Parag Kulkarni, "Design of Multi-agent System Architecture based on Association Mining for Cooperative Reinforcement Learning", *Spvryan's International Journal of Engineering Sciences & Technology (SEST)*, Volume 1, Issue 1, 2014.

[15] Dr. Hamid R. Berenji, David Vengerov "Learning, Cooperation, and Coordination in Multi-Agent Systems", *Intelligent Inference Systems Corp., Technical Report, October* 2000.

[16] Deepak A. Vidhate and Parag Kulkarni, "Performance enhancement of cooperative learning algorithms by improved decision-making for context-based application", *International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT) IEEE Xplorer*, pp 246-252, 2016

[17] Deepak A. Vidhate, Parag Kulkarni, "Enhanced Cooperative Multi-agent Learning Algorithms (ECMLA) using Reinforcement Learning", *International Conference on Computing, Analytics and Security Trends (CAST), IEEE Xplorer*, pp 556-561, 2017.