MULTI-ROBOT COORDINATION BY MACHINE LEARNING AND EVOLUTIONARY ALGORITHMS

Synopsis submitted
by
Arup Kumar Sadhu

Doctor of Philosophy (Engineering)

Artificial Intelligence Laboratory and Control Engineering Laboratory
Department of Electronics and Tele-Communication Engineering
Faculty Council of Engineering and Technology
Jadavpur University
Kolkata, India

2017

Synopsis of the Thesis entitled
MULTI-ROBOT COORDINATION BY MACHINE LEARNING AND EVOLUTIONARY ALGORITHMS

Submitted by **Arup Kumar Sadhu**
under the guidance of **Prof. Amit Konar,** ETCE Dept., JU

Coordination is a fundamental trait in lower level organisms as they used their collective effort to serve their goals. Hundreds of interesting examples of coordination are available in nature. For example, ants individually cannot carry a small food item, but they collectively carry quite a voluminous food to their nest. The tracing of the trajectory of motion of an ant following the pheromone deposited by its predecessor also is attractive. The queen bee in her nest directs the labor bees to specific directions by her dance patterns and gestures to collect food resources. These natural phenomena often remind us the scope of coordination among agents to utilize their collective intelligence and activities to serve complex goals.

Coordination and planning are closely related terminologies from the domain of multi-robot system. Planning refers to the collection of feasible steps required to reach a predefined goal from a given position. However, coordination indicates the skillful interaction among the agents to generate a feasible planning step. Therefore, coordination is an important issue in the field of multi-robot coordination to address complex real-world problems. Coordination usually is of three different types: cooperation, competition and mixed. As evident from their names, cooperation refers to improving the performance of the agents to serve complex goals, which otherwise seems to be very hard for an individual agent because of the restricted availability of hardware/software resources of the agents or deadline/energy limits of the tasks. Unlike cooperation, competition refers to serving conflicting goals by two (team of) agents. For example, in robot soccer, the two teams compete to win the game. Here, each team plans both offensively and defensively to score goals and thus act competitively. Mixed coordination indicates a mixture of cooperation and competition. In the example of a soccer game, inter-team competition and intra-team cooperation is the mixed coordination. Most of the common usage of coordination in robotics lies in cooperation of agents to serve a common goal. The thesis deals with the cooperation of robots/robotic agents to efficiently complete a complex task.

In recent times, researchers are taking keen interest to employ machine learning in multi-robot cooperation. The primary advantage of machine learning is to generate the action plans in sequence from the available sensory readings of the robots. In case of a single robot,

learning the action plans from the sensory readings is straight-forward. However, in the context of multi-robot, the positional changes of the other robots act as additional inputs for the learner robot, and thus learning is relatively difficult. Several machine learning and evolutionary algorithms have been adopted over the last two decades to handle the situations. The simplest of all is the supervised learning technique that requires an exhaustive list of sensory instances and the action plan by the robots. Usually, a human experimenter provides these data from his long acquaintance with such problems or by direct measurement of the sensory instances and decisions. The training instances being too large, sometimes has a negative influence to the engineer, and he/she feels it uncomfortable not to miss a single instance that carries valuable mapping from sensory instance to action plan by the robots.

Because of the difficulty of generating training instances and excessive computational overhead to learn those instances, coupled with the need for handling dynamic situations, researchers felt the importance of reinforcement learning (RL). In RL, we need not provide any training instance, but employ a critic who provides a feedback to the learning algorithm about the possible reward/penalty of the actions by the agent. The agent/s on receiving the approximate measure of penalty/reward understands which particular sensory-motor instances they need to learn for future planning applications. The dynamic nature of environment thus can easily be learned by RL. In the multi-agent scenario, RL needs to take care of learning in joint state/action space of the agents. Here, each agent learns the sensory-motor instances in the joint state/action space with an ultimate motive to learn the best actions for itself to optimize its rewards.

The superiority of evolutionary algorithms (EA) in optimizing diverse objective functions is subjected to the No Free Lunch Theorem (NFLT). According to NFLT, the expected effectiveness of any two traditional EAs across all possible optimization problems is identical. A self-evident implication of NFLT is that the elevated performance of one EA, say A, over the other, say B, for one class of optimization problems is counterbalanced by their respective performances over another class. It is therefore practically difficult to devise a universal EA that would solve all the problems. This apparently paves the way for hybridization of EAs with other optimization strategies, machine learning techniques, and heuristics.

In evolutionary computation paradigm, hybridization refers to the process of integrating the attractive features of two or more EAs synergistically to develop a new hybrid EA. The hybrid EA is expected to outperform its ancestors with respect to both accuracy and complexity over application-specific or general benchmark problems. The fusion of EAs through hybridization hence can be regarded as the key to overcome their individual limitations.

Hence, apart from the RL, hybridization of the evolutionary algorithms (EA) is also an effective approach to serve the purpose of multi-robot coordination in a complex environment. The primary objective of an EA in the context of multi-robot coordination is concerned with the minimization of the time consumed by the robots (i.e., the length of the path to be traversed by the robots) for complete traversal of the planned trajectory. In other words, robots plan their local trajectory, so that robots shifted from given positions to the next positions (sub-goals) in a time-optimal sense avoiding collision with the obstacles or the boundary of the world-map. The optimization algorithm is executed in each local planning step to move a small distance. Hence, cumulatively robots move to the desired goal position using the sequence of local planning. There are traces of literature on hybridization of the EAs.

Several algorithms for multi-agent learning are available in the literature, each with one specific flavor to optimize certain learning intents of the agents. Of these algorithms, quite a few interesting works on the MAQL have been reported in the literature. Among the state-of-the-art MAQL algorithms, the following need special mentions. Claus and Boutilier, aimed at solving the coordination problem using two types of reinforcement learners. The first one, called independent learner (IL), takes care of the learning behavior of individual agents by ignoring the presence of other agents. The second one, called joint action learner (JAL), considers all agents including the self to learn at joint action-space. Unlike JAL, in Team Q-learning proposed by Littman, an agent updates its Q-value at a joint state-action pair without utilizing associated agents' reward; rather the value function of the agent at the next joint state is evaluated by obtaining the maximum Q-value among the joint actions at the next joint state. Ville proposed Asymmetric-Q learning (AQL) algorithm, where the leader agents are capable of maintaining all the agents Q-tables. However, the follower agents are not allowed to maintain all the agents' Q-tables and hence, they just maximize their own rewards. In AQL, agents always achieve the pure strategy Nash equilibrium (NE), although there does exist mixed strategy NE. Hu and Wellman extended the Littman's Minimax Q-learning to the general-sum stochastic game (where the summation of all agents' payoff is neither zero nor constant) by taking into account of other agents' dynamics using NE. They also offered a proof of convergence of their algorithm. In case of multiple NE occurrences, one is selected optimally. Littman proposed Friend-or-Foe Q-learning (FQL) algorithm for general-sum games. In this algorithm, the learner is instructed to treat each other agent either as a friend in Friend Q-learning or as a foe in Foe Q-learning. Friend-or-Foe Q-learning provides a stronger convergence guarantee in comparison to that of the existing NE based learning rule. Greenwald and Hall proposed correlated Q-learning (CQL) employing correlated equilibrium (CE) to generalize both Nash Q-learning (NQL) and FQL. The bottlenecks of the above

MAQL algorithms are update policy selection for adaptation of the Q-tables in joint state-action space and the curse of dimensionality with an increase in the number of learning agents. Several attempts have been made to handle the curse of dimensionality in MAQL. Jelle and Nikos proposed Sparse Cooperative Q-learning, where a sparse representation of the joint state-action space of the agents is done by identifying the need for coordination among the agents at a joint state. Here, agents undertake coordination by their actions only in a few joint states. Hence, each agent maintains two Q-tables: one is the individual-action Q-table for un-coordinated joint states and another one is the joint action Q-table to represent the coordinated joint states. In case of uncoordinated states, a global Q-value is evaluated by adding the individual Q-values. Zinkevich offers a neural network based approach for generalized representation of the state-space for multi-agent coordination. By such generalization, agents (here robots) can avoid collision with an obstacle or other robots by collecting minimum information from the sensors. Reinaldo et al. proposed a novel algorithm to heuristically accelerate the TMAQL algorithms.

In the literature of MAQL agents either converge to NE or CE. The equilibrium-based MAQL algorithms are most popular for their inherent ability to determine optimal strategy (equilibrium) at a given joint state. Hu et al. identified the phenomenon of similar equilibria in different joint states and introduced the concept of equilibrium transfer to accelerate the state-of-the-art equilibrium-based MAQL (NQL and CQL). In equilibrium transfer, agents recycle the previously computed equilibria having very small transfer-loss. Recently Zhang et al. attempted to reduce the dimension of the Q-tables in NQL. The reduction is done by allowing the agents to store the Q-values in joint state-individual action space, instead of joint state-action space.

In the state-of-the-art MAQL (NQL and CQL), balancing exploration/exploitation during the learning phase is an important issue. Traditional approaches used to balance exploration/exploitation in MAQL are summarized here. The greedy exploration, although has wide publicity, needs to tune the value of which is time-costly. In the Boltzmann strategy, the action selection probability is controlled by tuning a control parameter (temperature) and by utilizing the Q-values due to all actions at a given state. Here, the setting of temperature to infinity (zero) implies pure exploration (exploitation). Unfortunately, the Boltzmann strategy antagonistically affects the speed of learning. Evolution of the Boltzmann strategy towards better performance is observed in a series of literature. However, the above selection mechanisms are not suitable for selecting a joint action preferred for the team (all the agents) because of the dissimilar joint Q-values offered by the agents at a common joint state-action pair. There are traces of literature concerning joint action selection at a joint state during

learning. However, with the best of our knowledge, there is no work in the literature, which considers the work, presented in this thesis.

The thesis includes six (6) chapters. Chapter 1 provides an introduction to the multi-robot coordination algorithms for complex real-world problems, including transportation of a box/stick, formation control for defense applications and soccer playing by multiple robots utilizing the principles of reinforcement learning, the theory of games, dynamic programming, and/or evolutionary algorithm. Naturally, this chapter provides a thorough survey of the existing literature of reinforcement learning with a brief overview of the evolutionary optimization to examine the role of the algorithms in the context of multi-agent coordination. Chapter 1 includes multi-robot coordination employing evolutionary optimization, and especially reinforcement learning for cooperative, competitive, and their composition for application to static and dynamic games. The latter part of the chapter deals with an overview of the metrics used to compare the performance of the algorithms while coordinating. Fundamental metrics for performance analysis are defined to study the learning and planning algorithms.

Chapter 2 offers learning-based planning algorithms, by extending the traditional multi-agent Q-learning algorithms (Nash Q-Learning and Correlated Q-Learning) for multi-robot coordination and planning. This extension is achieved by employing two interesting properties. The first property deals with the exploration of the team-goal (simultaneous success of all the robots) and the other property is related to the selection of joint action at a given joint state. The exploration of team-goal is realized by allowing the agents, capable of reaching their goals, to wait at their individual goal states, until remaining agents explore their individual goals synchronously or asynchronously. Selection of joint action, which is a crucial problem in traditional multi-agent Q-learning, is performed here by taking the intersection of individual preferred joint actions of all the agents. In case the resulting intersection is a null set, the individual actions are selected randomly or otherwise following classical techniques. The superiority of the proposed learning and learning-based planning algorithms are validated over contestant algorithms in terms of the speed of convergence and run-time complexity respectively.

In chapter 3, it is shown that robots may select the suboptimal equilibrium in presence of multiple types of equilibria (here Nash equilibrium or correlated equilibrium). In the above perspective, robots need to adapt to such a strategy, which can select the optimal equilibrium in each step of the learning and the planning. To address the bottleneck of the optimal equilibrium selection among multiple types, chapter 3 presents a novel consensus Q-learning for multi-robot coordination, by extending the equilibrium-based multi-agent Q-learning algorithms. It is also shown that a consensus (joint action) jointly satisfies the conditions of

the coordination type pure strategy Nash equilibrium and the pure strategy correlated equilibrium. The superiority of the proposed consensus Q-learning algorithm over traditional reference algorithms in terms of the average reward collection are shown in the experimental section. In addition, the proposed consensus-based planning algorithm is also verified considering the multi-robot stick-carrying problem as the testbed.

Unlike correlated Q-learning, Chapter 4 proposes an attractive approach to adapt composite rewards of all the agents in one Q-table in joint state-action space during learning, and subsequently, these rewards are employed to compute correlated equilibrium in the planning phase. Two separate models of multi-agent Q-learning have been proposed. If the success of only one agent is enough to make the team successful, then model-I is employed. However, if an agent's success is contingent upon other agents and simultaneous success of the agents is required then model-II is employed. It is also shown that the correlated equilibrium obtained by the proposed algorithms and by the traditional correlated Q-learning are identical. In order to restrict the exploration within the feasible joint states, constraint versions of the said algorithms are also proposed. Complexity analysis and experiments have been undertaken to validate the performance of the proposed algorithms in multi-robot planning on both simulated and real platforms.

Chapter 5 hybridizes the Firefly Algorithm and the Imperialist Competitive Algorithm. The above explained hybridization results in the Imperialist Competitive Firefly Algorithm, which is employed to determine the time-optimal trajectory of a stick, being carried by two robots, from a given starting position to a predefined goal position amidst static obstacles in a robot world-map. The motion dynamics of fireflies of the Firefly Algorithm is embedded into the socio-political evolution-based meta-heuristic Imperialist Competitive Algorithm. Also, the trade-off between the exploration and exploitation is balanced by modifying the random walk strategy based on the position of the candidate solutions in the search space. The superiority of the proposed Imperialist Competitive Firefly Algorithm is studied considering run-time and accuracy as the performance metrics. Finally, the proposed algorithm has been verified in a real-time multi-robot stick-carrying problem.

Chapter 6 concludes the thesis based on the analysis made, experimental and simulation results obtained from the earlier chapters. The chapter also examines the prospects of the thesis in view of the future research trends.

In summary, the thesis aimed at developing multi-robot coordination algorithms with a minimum computational burden and less storage requirement as compared to the traditional algorithms. The novelty, originality, and applicability of the thesis are illustrated below.

Chapter 1 introduces fundamentals of the multi-robot coordination. Chapter 2 offers two useful properties, which have been developed to speed-up the convergence of TMAQL

algorithms in view of the team-goal exploration, where team-goal exploration refers to the simultaneous exploration of individual goals. The first property accelerates exploration of the team-goal. Here, each agent accumulates high (immediate) reward for team-goal state-transition, thereby improving the entries in the Q-table for state-transitions leading to the team-goal. The Q-table thus obtained offers the team the additional benefit to identify the joint action leading to a transition to the team-goal during the planning, where TMAQL-based planning stops inadvertently. The second property directs an alternative approach to speed-up the convergence of TMAQL by identifying the preferred joint action for the team. Finding preferred joint action for the team is crucial when robots are acting synchronously in a tight cooperative system. The superiority of the proposed algorithms in Chapter 2 is verified both theoretically as well as experimentally in terms of the convergence speed and the run-time complexity.

Chapter 3 proposes the novel consensus Q-learning (CoQL), which addresses the equilibrium selection problem. In case multiple equilibria exist at a joint state by adapting the Q-functions at a consensus. Analytically it is shown that a consensus at a joint state is a coordination type pure strategy NE as well as a pure strategy CE. Experimentally, it is shown that the average rewards earned by the robots are more when adapting at consensus, than by either NE or CE.

Chapter 4 introduces a new dimension in the literature of the traditional CQL. In traditional CQL, CE is evaluated both in learning and planning phases. In Chapter 4, CE is computed partly in the learning and the rest in the planning phases, thereby requiring CE computation once only. It is shown in an analysis, that the CE obtained by the proposed techniques is same as that obtained by the traditional CQL algorithms. In addition, the computational cost to evaluate CE by the proposed techniques is much smaller than that obtained by traditional CQL algorithms for the following reasons. Computation of CE in the traditional CQL requires consulting m Q-tables in joint state-action space for m robots, whereas in the present context, we use a single Q-table in the joint state-action space for evaluation of CE. Complexity analysis (both time-and space-complexity) undertaken here confirms the last point. Two schemes are proposed: one for a loosely-and the other one for a tightly-coupled multi-robot system. Also, the problem-specific constraints are taken care of in Chapter 4 to avoid unwanted exploration of the infeasible state-space during the learning phase, thereby saving additional run-time complexity during the planning phase. Experiments are undertaken to validate the proposed concepts in simulated and practical multi-agent robotic platform (here Khepera-environment).

Chapter 5 offers the evolutionary optimization approach to address the multi-robot stick-carrying problem using the proposed Imperialist Competitive Firefly Algorithm (ICFA).

ICFA is the synergistic fusion of the motion dynamics of a firefly in the Firefly Algorithm (FA) and the local exploration capabilities of the Imperialist Competitive Algorithm. In ICA, an evolving colony is not guided by the experience of more powerful colonies within the same empire. However, in ICFA each colony attempts to contribute to the improvement of its governing empire by improving its socio-political attributes following the motion dynamics of a firefly in the FA. To improve the performance of the above mentioned hybrid algorithm further, the step-size for random movement of each firefly is modulated according to its relative position in the search space. An inferior solution is driven by the explorative force while a qualitative solution should be confined to its local neighborhood in the search space. The chapter also recommends a novel approach of evaluating the threshold value for uniting empires without imposing any serious computational overhead on the traditional ICA. Simulation and experimental results confirm the superiority of the proposed ICFA over the state-of-art techniques. Chapter 6 concludes the thesis with interesting future research directions.

Artificial Intelligence Laboratory and
Control Engineering Laboratory
Department of Electronics & Tele-Communication Engineering Arup Kumar Sadhu
Jadavpur University