Ad Hoc Teamwork by Learning Teammates' Task

(JAAMAS Extended Abstract)

Francisco S. Melo
INESC-ID/Instituto Superior Técnico
Universidade de Lisboa
Av. Prof. Dr. Cavaco Silva
2744-016 Porto Salvo, Portugal
fmelo@inesc-id.pt

Alberto Sardinha
INESC-ID/Instituto Superior Técnico
Universidade de Lisboa
Av. Prof. Dr. Cavaco Silva
2744-016 Porto Salvo, Portugal
jose.alberto.sardinha@tecnico.ulisboa.pt

ABSTRACT

We address ad hoc teamwork, where an agent must coordinate with other agents in an unknown common task without pre-defined coordination. We formalize the ad hoc teamwork problem as a sequential decision problem and propose (i) the use of an online learning approach that considers the different tasks depending on their ability to predict the behavior of the teammate; and (ii) a decision-theoretic approach that models the ad hoc teamwork problem as a partially observable Markov decision problem. We provide theoretical and empirical evaluation of the performance of our proposed methods in several domains of different complexity.

General Terms

Algorithms,Theory

Keywords

Ad hoc teamwork, online learning, POMDP

1. INTRODUCTION

The challenge of developing autonomous agents that are capable of cooperatively engaging with other unknown agents in a joint common task is known as the *ad hoc teamwork problem* [3]. In this paper, we break down the ad hoc teamwork problem into three challenges that an agent must address when paired within an ad hoc team. We claim that the agent may not always know beforehand the task that it is expected to complete. Therefore, it must *identify the target task*, *idenfy the teammates strategy* and *decide* accordingly.

Research on ad hoc teamwork mostly focused on the decision-making step, relying on different assumptions that simplify the other challenges. For example, Stone and Kraus [5] admit knowledge of teammate and task and use bandit algorithms to drive the teammate to the desired behavior. Posterior works no longer use teammate knowledge, assuming they are best responders, but still rely on task knowledge [4]. Recent works combine learning with planning, extending ad hoc teamwork to complex cooperative domains [1].

In this paper, we formalize the ad hoc teamwork problem as a sequential decision problem, in which the ad hoc agent does not know in advance the target task or how its teammates act towards completing it. Each task is represented as fully cooperative matrix game; the ad hoc agent identifies the target task by determining the corresponding payoff function from a set of possible alternatives. Teammates are assumed to follow a bounded-rationality best-response model [4], adapting their behavior to that of the learning agent. We the propose two novel approaches to the ad hoc teamwork problem: the first uses online learning to detect the target task and act accordingly; the second models the ad hoc teamwork problem as partially observable Markov decision process (POMDP) that can be solved using standard approaches from the literature. Finally, we present both theoretical bounds on and an empirical evaluation of the performance of our approaches in several domains.

Appears in: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016), J. Thangarajah, K. Tuyls, C. Jonker, S. Marsella (eds.), May 9–13, 2016, Singapore.*

2. THE AD HOC TEAMWORK PROBLEM

An agent placed in a team of unknown agents executing an unknown target task must, first of all, determine such target task. We assume that the target task τ^* lies in a known finite set $\mathcal{T}$ of tasks, each represented as a fully cooperative matrix game $(N, \times_{n=1}^{N}\mathcal{A}_n, u_\tau)$, where N is the number of agents, A_n is the action set of agent n and u_τ is the utility function of task τ. Once the task is identified, the agent must determine how the teammates address that task so that it can plan its own actions. These three steps (task identification, teammate identification and planning) should be expected from any ad hoc agent.

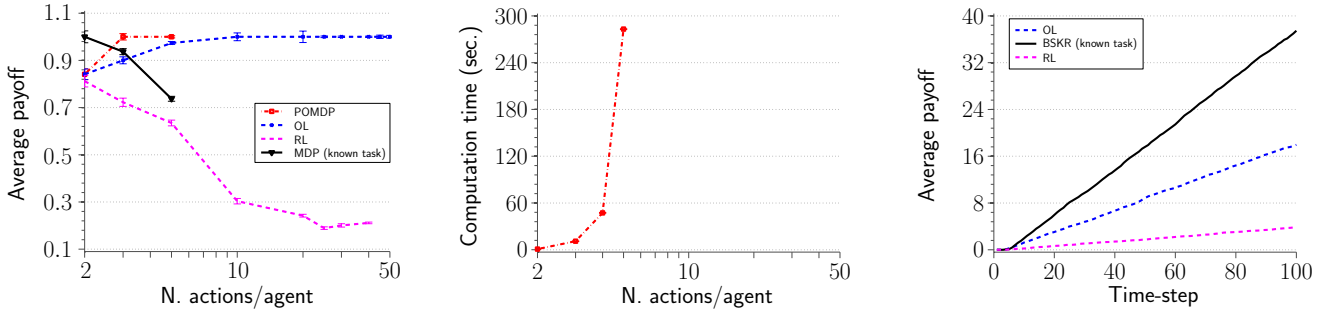
2.1 Online Learning (OL) Approach

We denote the ad hoc agent as α and its teammates as a single meta-agent $-\alpha$. As in other works [4], we assume $-\alpha$ is a best responder to the last M plays of α.

Our first approach requires the ad hoc agent to predict, at each time step n, the action $A_{-\alpha}(n)$ of its teammate. Therefore, at each step n, the ad hoc agent selects a joint action $\hat{A}(n) = \langle A_\alpha(n), \hat{A}_{-\alpha}(n)\rangle$ that includes its own action, $A_\alpha(n)$, and its prediction, $\hat{A}_{-\alpha}(n)$, and incurs a loss

$$\ell(\hat{A}(n), A_{-\alpha}(n)) = 1 - \delta(\hat{A}_{-\alpha}(n), A_{-\alpha}(n)). \qquad (1)$$

where δ is the Kroneker delta function. We cast the ad hoc teamwork problem as an *online prediction problem*, associating with each task $\tau \in \mathcal{T}$ an *expert* $E_\tau : \mathcal{H} \times \mathcal{A} \to [0, 1]$, where $E_\tau(h_{1:n}, a)$ is the probability of selecting the joint action a as a best response to the history $h_{1:n}$ according to



(a) (b) (c)

Figure 1: (a) Payoff in randomly generated scenarios vs the number of actions per agent. (b) POMDP computation time vs the number of actions per agent. (c) Payoff in the pursuit domain.

task τ. The cumulative loss of E_τ given the history $h_{1:n}$ is

$$L_\tau(h_{1:n}) \triangleq \sum_{t=0}^{n-1} \ell_\tau(h_{1:t}, a_{-\alpha}(t+1)),$$

where $a_{-\alpha}(t)$ is the actual action played by the teammate at time-step t after history $h_{1:t}$. Considering the *exponentially weighted average predictor* (EWAP) given by:

$$P(h_{1:n}, \hat{a}) \triangleq \frac{\sum_{\tau\in\mathcal{T}} e^{-\gamma_n L_\tau(h_{1:n})} E_\tau(h_{1:n}, \hat{a})}{\sum_{\tau\in\mathcal{T}} e^{-\gamma_n L_\tau(h_{1:n})}}, \qquad (2)$$

where γ_n is a positive parameter, we get the next result [2]:

THEOREM 1. *If $\gamma_t = \sqrt{2\ln|\mathcal{E}|/t^{1/2}}$ for all $t > 0$, then, for any finite history $h_{1:n} \in \mathcal{H}$, the* regret $R_n(P,\mathcal{E})$ *of an ad hoc agent following the EWAP is bounded above by $\sqrt{n/2\ln|\mathcal{E}|}$, where $R_n(P,\mathcal{E}) = \mathbb{E}\left[L_P(h_{1:n}) - L_\tau(h_{1:n})\right]$ and $L_P(h_{1:n})$ is the cumulative loss of predictor P at step n, given $h_{1:n}$.*

2.2 Decision-theoretic (POMDP) Approach

This section further considers prior knowledge about the target task, encoded as a prior distribution p_0, and the impact of the actions of agent on those of the teammate. The prior p_0 suggests a Bayesian approach to the ad hoc teamwork problem; the consideration of the agent α's actions impacts the way in which *regret* is defined.

We model the ad hoc teamwork problem as a POMDP $(\mathcal{X}, \mathcal{A}, \mathcal{Z}, \mathsf{P}, \mathsf{O}, c, \gamma)$, where at each time step n the state takes values in $\mathcal{X}$ and consists of the history of past M plays and the (unknown) target task. The action of the agent takes values in $\mathcal{A}$ and again includes its own action, $A_\alpha(n)$ and the prediction of the teammate's action, $\hat{A}_{-\alpha}(n)$. After selecting an action, the agent observes $A_{-\alpha}(n)$, which corresponds to the observation space $\mathcal{Z} = \mathcal{A}_{-\alpha}$. The *transition probabilities* P encode the state dynamics as a function of the agent's actions, and are factored in two components: the dynamics of the target task, which remains unchanged, and the dynamics of the history, which evolve according to the agent's actions and the best-response model of the teammate. Finally, the *observation probabilities*, O, map the current state to the teammate's most recent action, and the reward function r weights the payoff received from the interaction (determined by u_{τ^*}) with that of the loss in (1):

$$r(h,\tau,a) = \sum_{\hat{a},\hat{b}\in\mathcal{A}} E_\tau(h,\hat{a}) E_\tau(h,\hat{b})(K - \ell(\hat{a},a)) U_\tau(a_\alpha, \hat{b}_{-\alpha})$$

where K is a normalizing constant. We have modeled the ad hoc teamwork problem as a POMDP to which we can apply any POMDP solver. We have the following result [2]:

THEOREM 2. *The POMDP-based approach to the ad hoc teamwork problem is such that* $\lim_{n\to\infty} \frac{1}{n} R_n(P,\mathcal{E}) = 0$.

3. EMPIRICAL EVALUATION

We empirically evaluate the performance in random problems with increasing complexity (Figs. 1(a) and 1(b)) and the benchmark pursuit domain [1]. We compare the performance of the OL and POMDP approaches with the performance of an approach with task knowledge (MDP) and a standard RL approach. As seen in Fig. 1(a), the POMDP approach is clearly superior, although at a cost in computation (Fig. 1(b)). The two approaches proposed herein thus offer a delicate tradeoff between performance and complexity (see [2] for further discussion). In the pursuit domain (Fig. 1(c)), we compare our OL approach to the BSKR approach (from [1]), which uses full knowledge of the task. The domain is too large for the POMDP approach to be competitive (in computational terms). The OL approach is clearly superior to a naive RL approach, although not reaching the superior performance of BSKR. For a more detailed discussion, please refer to the full version of the paper [2].

Acknowledgements

This work was partially supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013 and the CMU-Portugal Program and its Information and Communications Technologies Institute, under project CMUP-ERI/HCI/0051/2013.

REFERENCES

[1] S. Barrett, P. Stone, S. Kraus, A. Rosenfeld. Teamwork with limited knowledge of teammates. *AAAI*, pp. 102-108, 2013.

[2] F. S. Melo, A. Sardinha. Ad hoc teamwork by learning teammates' task. *J. Autonomous Agents and Multiagent Systems*, 30(2):175-219, 2016.

[3] P. Stone, G. Kaminka, S. Kraus, J. Rosenschein. Ad hoc autonomous agent teams: Collaboration without pre-coordination. *AAAI*, pp. 1504-1509, 2010.

[4] P. Stone, G. Kaminka, J. Rosenschein. Leading a best-response teammate in an ad hoc team. *LNBIP, 59*, pp. 132-146, 2010.

[5] P. Stone and S. Kraus. To teach or not to teach?: Decision-making under uncertainty in ad hoc teams. *AAMAS*, pp. 117-124, 2010.