To appear in *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence Student Abstract and Poster Program* (AAAI), San Francisco, CA, August 2011.

Ad Hoc Teamwork in Variations of the Pursuit Domain

Samuel Barrett and **Peter Stone**
Dept. of Computer Science
The University of Texas at Austin
Austin, TX 78712 USA
(512)471-7316
{sbarrett,pstone}@cs.utexas.edu

Abstract

In multiagent team settings, the agents are often given a protocol for coordinating their actions. When such a protocol is not available, agents must engage in ad hoc teamwork to effectively cooperate with one another. A fully general ad hoc team agent needs to be capable of collaborating with a wide range of potential teammates on a varying set of joint tasks. This paper extends previous research in a new direction with the introduction of an efficient method for reasoning about the value of information. Then, we show how previous theoretical results can aid ad hoc agents in a set of testbed pursuit domains.

Introduction

As the number of autonomous agents in society grows, so does the need for them to interact with other agents effectively. Both robots and software agents are becoming more common in our society, and they are becoming more durable and robust, remaining deployed for increasing durations. Most existing methods for handling the interactions of agents require prior coordination, either in the form of coordination or communication protocols. However, as agents stay deployed for longer, new agents are likely to be introduced that may not share these protocols. Furthermore, a multitude of different agents are under development in different businesses and research laboratories. Unfortunately, it is unlikely that these agents will all share a communication protocol. Therefore, it is desirable for agents to be capable of adapting to new agents and learning to cooperate with previously unseen agents as part of an *ad hoc* team.

In a recent AAAI challenge paper, Stone et al. (2010) introduced the concept of an *ad hoc team setting*, specifying it as a problem in which team coordination strategies cannot be specified a priori. They further presented a framework for evaluating the performance of an ad hoc team agent with respect to a domain and a set of possible teammates. In an ad hoc team, agents need to cooperate with previously unseen teammates. Rather than developing protocols for coordinating an entire team, ad hoc team research focuses on developing agents that cooperate with teammates in the absence of such explicit protocols. Therefore, we consider a single agent cooperating with teammates that may or may not adapt to its behavior. For this work, we adopt essentially the same evaluation framework proposed by Stone et al. (2010), where the performance of the ad hoc team agent depends on the distribution of problem domains and the distribution of possible teammates.

This paper investigates several empirical ad hoc teamwork scenarios, and proposes a new algorithm for efficiently planning while considering the value of information for determining teammates' behaviors. In addition, it shows how existing theoretical results can be applied to empirical problems. This work unifies current research on ad hoc teams by applying existing results to variations of the same problem.

Ad Hoc Teamwork in the Pursuit Domain

The pursuit domain has become a popular setting for multiagent research (Stone and Veloso 2000) since it was introduced by Benda et al. (1986). It lends itself well to ad hoc team problems as it requires multiple predators to cooperate to capture the prey. Two versions of the pursuit domain have been investigated using an agent that plans efficiently using Monte Carlo Tree Search (MCTS) by Barrett et al. (2011). To handle uncertainty about its teammates, the ad hoc agent updates the relative probabilities of the known behaviors. At each time step, the agent plans using MCTS and samples from the possible teammate models with respect to their relative probabilities. However, actions that the ad hoc agent takes can reveal which behavior its teammates are playing, and this information may boost future performance.

The Value of Information

Reasoning about the value of information can allow an ad hoc agent to optimally handle the trade-off between exploring its teammates' behavior and exploiting its current models. Considering the value of information is not a new idea (Howard 1966), but computation cost of this reasoning can be high (Dearden, Friedman, and Andre 1999; Ross, Chaib-draa, and Pineau 2008). Therefore, we propose one of the main contributions of this paper: an efficient algorithm for approximating the value of information when handling unknown behaviors for teammates. Rather than searching through the space of world states, the agent includes its beliefs about its teammates' behaviors in the state, resulting in a world belief state. For this problem, these beliefs represent the relative probabilities of the different known behaviors that the teammates may be playing. Using MCTS over this world belief state implicitly calculates the

value of information because if an action results in a world belief state that leads to a higher expected reward, the agent will prefer this action. We keep an additional estimate of the state-action values for the original state, and combine these two estimates; using the original state-action values to bias the agent's exploration towards actions that work well on the world state, regardless of the belief state.

The ad hoc team agent is given a set of possible behaviors that its teammates may be following, and by observing its teammates over time, it can identify which behavior they are following. Since some predators may react badly to poor early moves by the ad hoc team agent, it is vital for the ad hoc agent to quickly identify the correct model of its teammates and only deviate from what's expected when there is a large advantage to be gained. Therefore, by reasoning about the value of information, the ad hoc agent can take actions that help it identify the correct behavior before its teammates detect deviations in its behavior. The results from our tests suggest that reasoning about the value of information can be done efficiently and can significantly improve results.

Repeated Interactions with a Best Response Agent

While previous work assumes that the ad hoc agent interacts with its teammates for only one episode, many teamwork settings allow for multiple interactions among the same teammates. In this case, long-term learning (across episodes) is both possible and very useful. We assume that both predators choose to play a high level behavior until the prey is captured rather than selecting directions to move at each time step and that the teammate chooses behaviors by best response. If both predators know how each pair of behaviors performs, this problem can be modeled as a repeated normal-form game in which the agents share the payoffs. In this setting, there is a matrix of shared payoffs and two agents, and there is one cell in the payoff matrix with the highest reward that is best for both agents. However, if the ad hoc agent jumps immediately to the corresponding behavior, it may incur a high loss before the best response agent moves to the best action, where the loss is defined as the difference between the maximum possible reward and the received reward. Therefore, it may be desirable for the ad hoc agent to take a longer path through the payoffs, minimizing the losses. Stone et al. (2010) investigated exactly the class of ad hoc teamwork problems that can be modeled by this normal-form game formulation, and their results aid in the development of ad hoc team agents for this empirical domain.

Teaching a Novice Agent

To this point, we have assumed that the teammate has complete knowledge about the performance of the behaviors, but in some settings this is not the case. To investigate such settings, we introduce a version of the pursuit domain in which the teammate starts with no knowledge about each behavior and must explore the behaviors to estimate their performance. In this version, two predators take turns trying to capture the prey, but they do not directly interact although they can observe the results of the other's actions. Therefore, a single predator must capture the prey independently by occupying the same cell as the prey, but the two predators are a team and share rewards. The ad hoc team agent has full knowledge about the performance of the behaviors, while its teammate starts with no knowledge and acts greedily with respect to the behaviors' observed sample means. However, the teammate is not able to execute all of the behaviors that are available to the ad hoc agent, including the one with the best expected reward. Therefore, there is a cost to the ad hoc team agent foregoing this best behavior in favor of playing another that will teach its teammate. Close examination of this problem reveals that it can be modeled by a multi-armed bandit (MAB), such as that proposed by Stone and Kraus (2010), where the different behaviors correspond to different arms of the bandit. Stone and Kraus's theoretical results apply to this domain and aid in the development of an effective ad hoc team agent.

Conclusion

This paper introduces an efficient algorithm for reasoning about the value of information and shows that this reasoning can significantly aid an agent dealing with uncertainty about its teammates' behavior. We then explore variations of the pursuit domain and reveal how existing theoretical results apply to these formulations. Doing so brings together several current ad hoc team research results by applying them all to problems in the pursuit domain.

Acknowledgments

This work has taken place in the Learning Agents Research Group (LARG) at the Artificial Intelligence Laboratory, The University of Texas at Austin. LARG research is supported in part by grants from the National Science Foundation (IIS-0917122), ONR (N00014-09-1-0658), and the Federal Highway Administration (DTFH61-07-H-00030). Samuel Barrett is supported by an NDSEG fellowship.

References

Barrett, S.; Stone, P.; and Kraus, S. 2011. Empirical evaluation of ad hoc teamwork in the pursuit domain. In *AAMAS '11*. To appear.

Benda, M.; Jagannathan, V.; and Dodhiawala, R. 1986. On optimal cooperation of knowledge sources - an empirical investigation. Technical Report BCS–G2010–28, Boeing Advanced Technology Center, Boeing Computing Services.

Dearden, R.; Friedman, N.; and Andre, D. 1999. Model-based Bayesian exploration. In *UAI*, 150–15.

Howard, R. 1966. Information value theory. *Systems Science and Cybernetics, IEEE Transactions on* 2(1):22 –26.

Ross, S.; Chaib-draa, B.; and Pineau, J. 2008. Bayesian reinforcement learning in continuous POMDPs with application to robot navigation. In *ICRA*, 2845–2851.

Stone, P., and Kraus, S. 2010. To teach or not to teach? Decision making under uncertainty in ad hoc teams. In *AAMAS '10*.

Stone, P., and Veloso, M. 2000. Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots* 8(3):345–383.

Stone, P.; Kaminka, G. A.; Kraus, S.; and Rosenschein, J. S. 2010. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *AAAI '10*.

Stone, P.; Kaminka, G. A.; and Rosenschein, J. S. 2010. Leading a best-response teammate in an ad hoc team. In *AMEC*.