Exploiting Structure and Utilizing Agent-Centric Rewards to Promote Coordination in Large Multiagent Systems

(Extended Abstract)

Chris HolmesParker
Oregon State University
442 Rogers Hall
Corvallis, OR 97331
holmespc@onid.orst.edu

Adrian Agogino
USCS at NASA Ames
Mail Stop 269-3
Moffett Field, CA 94035
adrian.k.agogino@nasa.gov

Kagan Tumer
Oregon State University
204 Rogers Hall
Corvallis, OR 97331
kagan.tumer@oregonstate.edu

ABSTRACT

A goal within the field of multiagent systems is to achieve scaling to large systems involving hundreds or thousands of agents. In such systems the communication requirements for agents as well as the individual agents' ability to make decisions both play critical roles in performance. We take an incremental step towards improving scalability in such systems by introducing a novel algorithm that conglomerates three well-known existing techniques to address both agent communication requirements as well as decision making within large multiagent systems. In particular, we couple a Factored-Action Factored Markov Decision Process (FA-FMDP) framework which exploits problem structure and establishes localized rewards for agents (reducing communication requirements) with reinforcement learning using agent-centric difference rewards which addresses agent decision making and promotes coordination by addressing the structural credit assignment problem. We demonstrate our algorithms performance compared to two other popular reward techniques (global, local) with up to 10,000 agents.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Experimentation

Keywords

Factored MDPs, Reward Shaping, Scalability

1. INTRODUCTION

Controlling large systems of autonomous agents represents a complex coordination problem. Here, agents must learn a set of joint-actions such that they optimize collective objective. In order to be effective, solutions must address two key issues present within such systems: i) communication requirements (complete communication within large multiagent systems is frequently infeasible [4]), and ii) decision making (agents must be able to learn to coordinate their joint actions [4]). One way of reducing the coordination costs of individual agents involves exploiting the problem structure and decomposing the system [4]. This work focuses on decomposing a large MDP with a system reward into several smaller joint MDPs or Factored MDPs with localized rewards [3, 5]. Such a decomposition is not the same as solving a disjoint set of smaller MDPs in parallel. On the contrary, in many FMDPs (such as the FA-FMDPs used in this work) significant coordination is required between all agents within the system due to the heavy coupling between their localized rewards. Thus, solving for joint-actions in such cases represents a formidable coordination task for the agents. We are interested specifically in large multiagent systems in which after the system has been decomposed, the coordination problem is still exceedingly difficult.

We present a novel algorithm that combines a FA-FMDP framework, reinforcement learning, and agent-specific shaped rewards (difference rewards) in order to promote coordination and scaling in large multiagent systems. This algorithm addresses three key issues present within these systems: i) it reduces the per-agent communication and coordination requirements by exploiting problem structure, ii) it enables agent decision making by using reinforcement learning, and iii) it addresses the structural credit assignment problem by using agent-specific shaped rewards (difference rewards).

Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013)*, Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.

2. BACKGROUND

Factored Markov Decision Processes (FMDPs) represent one approach to generalization, which exploits the underlying structure of the problem in multiagent systems. The idea of representing a large MDP using a factored model was first proposed by Boutilier et al. [2]. A factored MDP is a mathematical framework for sequential decision problems in stochastic domains [3]. It thus provides an underlying semantics for the task of planning under uncertainty [3]. Factored MDPs (FMDPs) are a representation language that allows us to exploit problem structure to represent exponentially large MDPs very compactly [3]. We focus on Factored-Action Factored Markov Decision Processes (FA-FMDPs), however our algorithm can easily be extended into Factored-State Factored Markov Decision Processes (FS-FMDPs) or Factored Markov Decision Processes which are factored with respect to both states and actions (FMDPs).

2.1 Multi-Target Defect Combination Problem

This problem assumes that there exists a set of imperfect sensors, $\mathbf{X}$, which have constant attenuations due to manufacturing defects or imperfections. Each of the sensors, x_j, has an associated attenuation, ζ_j, (which can be positive or negative) in its reading, such that if it is taking a measurement of A (actual value) it measures $A + \zeta_j$ where ζ_j is the device's individual error. The problem then becomes how to best choose a subset of the $\mathbf{X}$ sensors that minimizes the aggregated attenuation over the set of targets within the domain:

$$R = \sum_{m=1}^{M} \frac{\left| \sum_{j=1}^{N} n_{j,m} \zeta_j \right|}{\sum_{j=1}^{N} n_{j,m}} \quad (1)$$

where R is the aggregated attenuation (system reward) of the combined sensor readings over the set of M targets, ζ_j is the attenuation of a particular sensor j, N is the number of sensors, and $n_j \in \{0, 1\}$ is an indicator based upon whether sensor j chooses to be "on" or "off". In this setting, each sensing agent, j', is able to observe $C_{j'}$ of the M targets in the system simultaneously. Each individual agent's reward function, r_j, becomes the sum of the attenuations of the targets it senses as follows:

$$r_{j'} = \sum_{i \in C_{j'}} \frac{\left| \sum_{j \in \Gamma(i)} n_j \zeta_j \right|}{\sum_{j \in \Gamma(i)} n_j} \quad (2)$$

where $r_{j'}$ is the reward of agent j', $i \in C_{j'}$ is the set of $R_i \in C_{j'}$ that agent j' effects, $j \in \Gamma(i)$ is the subset of agents that are impacting reward R_i, ζ_j is the attenuation of agent j's sensor, and n_j is an indicator function that takes the value $\{0, 1\}$ that determines whether sensor j chose to be "on" or "off". We now derive the difference rewards for agents within the MTDCP domain ($D_{j'} = r_{j'} - r_{j', -z_{j'}}$ [1]):

$$D_{j'} = \sum_{i \in C_{j'}} \frac{\left| \sum_{j \in \Gamma(i)} n_j \zeta_j \right|}{\sum_{j \in \Gamma(i)} n_j} - \sum_{i \in C_{j'}} \frac{\left| \sum_{j \in \Gamma(i), j \neq j'} n_j \zeta_j \right|}{\sum_{j \in \Gamma(i), j \neq j'} n_j} \quad (3)$$

where D_j is the difference reward for agent j, $r_{j'}$ is the reward of agent j' which depends upon all agents involved in $R_i \forall i \in C_{j'}$, and $r_{j', -z_{-j'}}$ is agent j''s reward without the contribution of agent j'.

3. RESULTS

Agents using a random policy, S, perform poorly, as they fail to coordinate their actions intelligently. Agents using traditional global rewards, R, also perform poorly because agents struggle to coordinate their actions due to noise on their learning signals caused by the actions of other agents. Agents using localized rewards r_j defined by the FA-FMDP structure perform marginally better, but are still unable to coordinate well. Agents using difference rewards D and estimated difference rewards EDR within an FA-FMDP framework achieve good performance. This is because the FA-FMDP reduces the per-agent coordination complexity for individual agents into localized settings while difference rewards address the structural credit assignment problem.

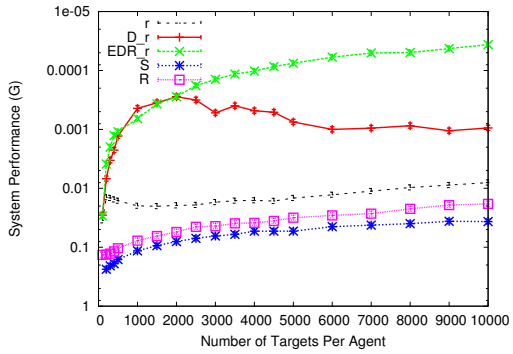

Figure 1: 10,000 agents with 100 targets and each agent is assigned to 10 targets. Agents using our algorithm D_r and EDR_r outperform all other methods because it reduces the communication requirements for agents while simultaneously addressing the structural credit assignment problem.

4. DISCUSSION

In this work, we showed that in many large multiagent systems the coordination requirements for agents are overwhelming, resulting in poor coordination and performance. To address this, we introduced a novel algorithm which combined three existing techniques used within the multiagent literature. Our algorithm reduced communication requirements for individual agents and defined localized rewards by structuring the system using an FA-FMDP framework and coupled reinforcement learning with agent-centric difference rewards in order to significantly improve scalability and performance within large multiagent systems. Our algorithm combined the benefits of an FA-FMDP structure (localized rewards, reduced communication requirements) with reinforcement learning using difference rewards which readily addressed the structural credit assignment problem, promoting coordination and good decision making. Our algorithm significantly outperformed agents using global rewards as well as localized rewards defined by the FA-FMDP framework for scaling up to 10,000 agents in two domains.

5. REFERENCES

[1] A. Agogino, C. HolmesParker, and K. Tumer. Evolving large scale uav communication system. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, Philadelphia, PA, July 2012.

[2] C. Boutilier, T. Dean, and S. Hanks. Decision-theoretic planning - structural assumptions and computational leverage. *Journal of Artificial Intelligence Research (JAIR)*, 1999.

[3] C. Guestrin, D. Koller, R. Parr, and S. Venkataraman. Efficient solution algorithms for factored mdps. *Journal of Artificial Intelligence Research (JAIR)*, 2003.

[4] L. Panait and S. Luke. Cooperative multi-agent learning - the state of the art. *Journal of Autonomous Agents and MultiAgent Systems (JAAMAS)*, 2005.

[5] A. Strehl, C. Diuk, and M. Littman. Efficient structure learning in factored-state mdps. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2007.