Decentralized Coordination via Task Decomposition and Reward Shaping

(Extended Abstract)

Atil Iscen
Oregon State University
Corvallis, OR, 97331, USA
iscena@onid.orst.edu

Kagan Tumer
Oregon State University
Corvallis, OR, 97331, USA
kagan.tumer@oregonstate.edu

ABSTRACT

In this work, we introduce a method for decentralized coordination in cooperative multiagent multi-task problems where the subtasks and agents are homogeneous. Using the method proposed, the agents cooperate at the high level task selection using the knowledge they gather by learning subtasks. We introduce a subtask selection method for single agent multi-task MDPs and we extend the work to multiagent multi-task MDPs by using reward shaping at the subtask level to coordinate the agents. Our results on a multi-rover problem show that agents which use the combination of task decomposition and subtask based difference rewards result in significant improvement both in terms of learning speed, and converged policies.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence

General Terms

Algorithms

Keywords

Reinforcement Learning, Reward Shaping, Cooperation

1. INTRODUCTION

Multi-task problems have a repetitive structure that allows a decomposition into smaller subtasks, each of which can be solved significantly more easily than the full problem. In this paper we address multiagent problems with homogeneous subtasks that are additive and utility independent. The reward provided by the MDP is a sum of the rewards for all the subtasks to which the MDP is decomposed. This problem has generally been addressed from a single learning agent perspective. However, when multiple agents are cooperating on a multitask problem, combining the multitask decomposition problem with multiagent learning brings a new dimension: subtask sharing between agents.

In this work, we approach the multiagent multi-task problem from a decentralized perspective. Task decomposition still takes place, but it is done at the agent level. Each agent decomposes the problem and decides which subtask to select. Clearly, this problem now requires coordination of the agents on selecting the subtasks. To establish cooperation, we decompose the problem into multiagent single task problems first, and use reward shaping at the subtask level. That way, the task selection is based on the agents competence on the tasks. Since lower level performance is based on the shaped reward, higher level selection

2. BACKGROUND

Reinforcement Learning (RL) is a learning method, where an agent learns from its experiences with an environment [4]. In RL, the agent is expected to learn the optimal behavior using the rewards given by the environment as feedback. At a given timestep, the agent gets the state information from the environment, then decides on an action and according to the state and the action of the agent, the environment returns the new state and reward. The Q value of a state action pair, denoted by $Q_\pi(s,a)$, is the estimation of the sum of all the rewards that one agent would get by taking action a at state s and following policy π. In RL algorithms such as Q learning or Sarsa, through the learning process, the agent improves the estimation of Q values and the policy to increase performance for the given problem.

For problems that contains multiple tasks, task decomposition is a commonly studied method. One way to represent task decomposition is using weakly coupled MDPs where decomposed MDPs have minimal amount of connection [2]. For multiagent learning with multiple tasks, assignment based task decomposition is previously used to decompose the problem into smaller problems containing an agent and assigned subtask [3]. In this approach, the agent subtask assignments are made by a centralized mechanism.

3. DECENTRALIZED COORDINATION

In a typical RL algorithm, the decisions are based on Q values that are higher for states closer to the goal state. Using this principle, we propose a two layer learning mechanism where top layer decision is based on the Q values of the lower level subtask learning. For a problem containing multiple subtasks, if the agent uses flat learning, the agent's Q value will be propagated from the closest goal state. On the other hand, in a learning with task decomposition, since each subtask is handled independently, Q values are inde-

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.

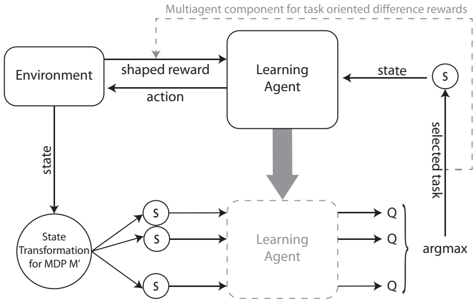

Figure 1: The learning diagram of HELM.

pendent for different tasks. The state given to the agent is decomposed into different states for different tasks. The learning agent is able to check the Q values of these decomposed states. When the learning for a task is converged to the optimal policy, these Q values for each task will be higher depending on how close the agent is to succeeding that specific task.

Since the agent now has the ability to evaluate current state for each of the tasks, the decision of which task to choose becomes straightforward: the maximum one. The upper level decision becomes selecting the maximum of the potential values for each task and lower level learning becomes a flat model with an MDP containing single task. On the other hand, since the primitive action selected by the agent is towards one of the subtasks, using reward signal given by the environment can be confusing for the agent. Since we decomposed the task, the agent uses the reward of the chosen subtask. This concept can also be considered as reward shaping with task decomposition.

Figure 1 illustrates this process. Formally, if we restate the idea described, at every timestep, the agent:

- Takes state s in MDP M containing many tasks. Takes reward r and update Q values according to the previously selected task and action.
- Decomposes into subtasks with single task MDP M'. For every task i, transforms the state s into state s'_i
- Selects task that maximizes Q value and selects an action according to the selected task.

The general idea is evaluation of current state for each of the subtasks and selecting the subtask that gives the highest Q value in M'. We call this method as **H**igh level **E**valuation of **L**ow level **M**DP using Q values (**HELM**).

For multiagent multi-task problems, first intuition is to decompose the problem into single agent single task problems. Since this approach does not provide any coordination, each agent would pick the easiest task ignoring the cooperation necessity of the problem. Instead, we decompose the problem to multiagent single-task problems. For every decomposition, the agents still consider all the agents of the environment. The cooperation is provided by introducing task oriented difference rewards as a reward shaping method. Difference rewards is a previously defined reward shaping method that provides individual rewards to each team member based on their contribution to the team to promote cooperative behavior [1]. By integrating Difference Rewards at the subtask level, the Q values are updated while taking cooperation into account. Since in HELM, the agents estimate their competence on each task and pick the best one according to the Q values, low level difference rewards provide coordination at the high level selection.

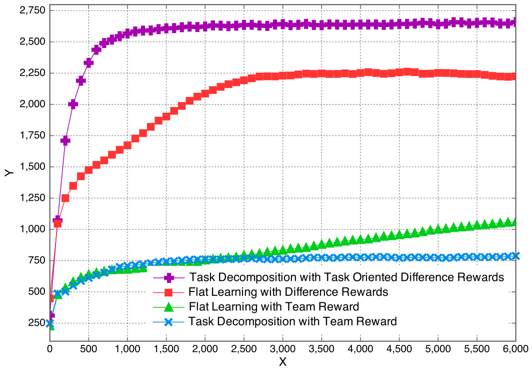

Figure 2: Results of task decomposition and reward shaping in continuous rover domain with 12 agents and 12 POIs

The algorithm is tested on a rover domain where multiple rovers learn to observe multiple Points of Interests (POI). The goal is to increase total amount of observation and the amount of observation for each POI is inversely proportional to the distance of the closest agent to that POI. The results show that, task decomposition and reward shaping together, results in a faster learning speed and a better converged policy (Figure 2).

4. CONCLUSIONS AND FUTURE WORK

We introduced a decentralized task decomposition algorithm for multiagent multi-task problems. The selection of the different tasks are evaluated through Q values of the decomposed tasks, and we integrated reward shaping to the method by introducing a task oriented version of the difference rewards. The method proposed is studied on the Rover Domain and experimental results show significant improvement on converged policy and significant decrease on time to learn. As a future research, we are working on automated discovery of the decomposition function.

Acknowledgements: This work was partially supported by NSF grants CNS-0931591 and CMMI-0928076.

5. REFERENCES

[1] A. K. Agogino and K. Tumer. Analyzing and visualizing multiagent rewards in dynamic and stochastic domains. *Autonomous Agents and Multi-Agent Systems*, 17:320–338, October 2008.

[2] N. Meuleau, M. Hauskrecht, K. eung Kim, L. Peshkin, L. P. Kaelbling, and T. Dean. Solving very large weakly coupled markov decision processes. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 165–172, 1998.

[3] S. Proper and P. Tadepalli. Solving multiagent assignment markov decision processes. AAMAS '09, pages 681–688, Richland, SC, 2009.

[4] R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning.* MIT Press, Cambridge, MA, USA, 1st edition, 1998.