Learning in Multi-Robot Systems

Maja J Matarić
Volen Center for Complex Systems
Computer Science Department
Brandeis University
Waltham, MA 02254
maja@cs.brandeis.edu

Learning in Situated Domains

Introduction

Reinforcement learning (RL) has been successfully applied to a variety of domains, and has recently been attempted on situated agents such as mobile robots. While simulation results are encouraging, work on physical robots has been slow to repeated that success. The key challenges of situated domains include: 1) modeling a combination of discrete and continuous state spaces based on multimodal perceptual inputs; 2) modeling real-world events that may neither be caused directly by the agents nor perceived by it, but subsequently affect its behavior; 3) the number of learning trials reasonably available to an agent and the non-uniform exploration of the learning space mandated by the agent's external environment; 4) dealing with multiple concurrent and sequential goals; 5) modeling a combination of discrete and continuous, immediate and delayed, multimodal feedback that may be available to the agent.

Designing Reward Functions

Rather than encode knowledge explicitly, RL methods hide it in the reinforcement function which often employs some *ad hoc* embedding of the domain semantics. One more direct way to utilize implicit domain knowledge is to convert reward functions into error signals, akin to those used in learning control. Immediate reinforcement in RL is a weak version of error signals, using only the sign of the error but not the magnitude. Intermittent reinforcement can be used similarly, by weighting the reward according to the accomplished progress.

We suggest that such reinforcement can be introduced 1) by reinforcing multiple goals, and 2) by using progress estimators. Since situated agents have multiple goals, it is straightforward to reinforce each one individually, with a *heterogeneous reinforcement function*, rather than to attempt to collapse them into a monolithic goal function. However, multiple goals are not sufficient for speeding up situated learning if each of them involves a complex sequence of actions. Such time-delayed goals are aided by progress metrics along the way, in addition to reinforcement upon achievement. We propose *progress estimators*, functions which provide positive or negative

Figure 1: R2 robots used to learn foraging.

reinforcement based on immediate measurable progress relative to specific goals. These "partial internal critics" serve a number of important functions in noisy worlds: they decrease the learner's sensitivity to intermittent errors, they encourage exploration and minimize thrashing, and they decrease the probability of fortuitous rewards for inappropriate behavior that happened, by chance, to achieve the desired goal. For a detailed discussion please see Matarić (1994).

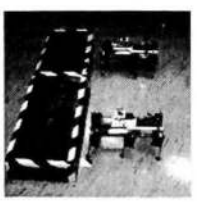

Figure 2: Genghis-II six-legged robots used to learn box-pushing.

Experimental Design

Both of our learning experiments were conducted on fully autonomous mobile robots on-board power, sensing, and computation. The first set of experiments was done with 4 IS Robotics R2 robots equipped with bump and infra-red sensors for detecting collisions and contacts, radio transceivers for positioning, communication, and data gathering, and situated on a differentially steerable wheeled base equipped with a gripper (Figure 1). The second set of experiments was done on 2 IS Robotics Genghis II robots, equipped with two whisker contact

sensors, an array of 5 pyro-electric sensor for detecting the location of the goal (the light), and using six-legged alternating tripod gate for propulsion (Figure 2). All of the robots were programmed in the Behavior Language and were controlled by collections of parallel, concurrently active behaviors that gather sensory information, drive effectors, monitor progress, and contribute reinforcement.

The Learning Tasks

The first learning task consisted of finding a mapping between conditions and behaviors into an efficient policy for group foraging. Foraging was chosen because it is a nontrivial and biologically inspired task, and because our previous group behavior work (Matarić 1992) provided the basis behavior repertoire from which to learn behavior selection, consisting of *avoiding*, *dispersing*, *searching*, *homing*, and *resting*. Utility behaviors for grasping and dropping were hard-wired as were their conditions. By considering only the space of conditions necessary and sufficient for triggering the above behaviors, the agents' learning space was reduced to the power set of the following state variables: *have-puck?*, *at-home?*, *near-intruder?*, and *night-time?*. The reduced foraging task should, in theory, be easily learnable. In practice, however, quick and uniform exploration is not possible in the noisy multi-agent domain.

The second learning task consisted of finding a policy for each of the robots to cooperatively push a long box to the goal. Unlike the foraging task, box-pushing required careful coordination between the agents, in turn requiring either accurate sensing, or communication, or both. The task is designed so that a single-agent solution, due to the size and shape of the box, is much less efficient than an effective two-agent solution, but the two-agent solution requires intricate cooperation or the box is pushed in the wrong direction or out of reach of one of the robots. The task was decomposed into basic behaviors: *pushing*, *pausing*, *turning*. The task required that each agent learn not only its own strategy for keeping the box within reach and moving toward the goal, but also the right behaviors in response to the other agent, as sensed through the state of the box and as communicated between the agents. The details of the experiments and the data are described in Simsarian & Matarić (1995).

Learning Results

The reinforcement learning algorithms we used summed all of the multimodal reinforcement over time. Behaviors were switched based on external events, as well as inputs from internal progress estimators. Reinforcement was based on a collection of internal functions that monitored external events and internal progress estimators. Learning was continuous and incremental over the lifetime of the agent.

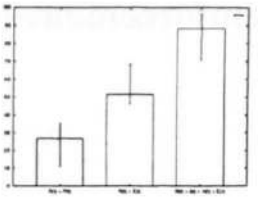

Figure 3: The performance of the three reinforcement strategies on learning to forage. The x-axis shows the three reinforcement strategies. The y-axis maps the percent of the correct policy the agents learned, averaged over twenty trials.

Both learning experiments were evaluated first by comparing the system performance to the control hard-wired behavior for foraging and for box-pushing. Second, the foraging learning performance was also compared to two alternative approaches, one using only multimodal reinforcement but no progress estimators, and the other using traditional Q-learning with positive reinforcement when a puck was dropped in the home region (Figure 3).

Summary

The goal of this work has been to bring to light some of the important properties of situated domains, and their impact on the existing reinforcement learning strategies. We have argued that the noisy and inconsistent properties of complex worlds require the use of domain knowledge. We proposed a principled approach to embedding such knowledge into the reinforcement based on utilizing heterogeneous reward functions and goal-specific progress estimators. We believe that these strategies take advantage of the information readily available to situated agents, make learning possible in complex dynamic worlds, and accelerate it in any domain.

References

Matarić, M. J. (1992), Designing Emergent Behaviors: From Local Interactions to Collective Intelligence, *in* J.-A. Meyer, H. Roitblat & S. Wilson, eds, 'From Animals to Animats: International Conference on Simulation of Adaptive Behavior'

Matarić, M. J. (1994), Interaction and Intelligent Behavior, Technical Report AI-TR-1495, MIT Artificial Intelligence Lab.

Simsarian, K. T. & Matarić, M. J. (1995), Learning to Cooperate Using Two Six-Legged Mobile Robots, *in* 'Proceedings, Third European Workshop of Learning Robots', Heraklion, Crete, Greece.