Efficient Apprenticeship Learning with Smart Humans

Kaushik Subramanian
Rutgers University
Dept. of Electrical and Computer Engineering
Piscataway, NJ
kausubbu@eden.rutgers.edu

Michael L. Littman
Rutgers University
Dept. of Computer Science
Piscataway, NJ
mlittman@cs.rutgers.edu

Abstract

This report describes a generalized apprenticeship learning protocol for reinforcement-learning agents with access to a teacher. The teacher interacts with the agent by providing policy traces (transition and reward observations). We characterize sufficient conditions of the underlying models for efficient apprenticeship learning and link this criteria to two established learnability classes (KWIK and Mistake Bound). We demonstrate our approach in a conjunctive learning task that would be too slow to learn in the autonomous setting. We show that the agent can guarantee near-optimal performance with only a polynomial number of examples from a human teacher and can efficiently learn in real world environments with sensor imprecision and stochasticity.

Introduction

Learning by Demonstration (LbD) is a powerful technique for teaching a robot new behaviors without extensive programming. A human, playing the role of the teacher, has the task of effectively communicating the required behavior to the artificial learning agent through various methods of interaction. An example of such a method is the apprenticeship learning protocol (Abbeel and Ng 2005) for Reinforcement-Learning (RL) agents that attempts to learn the dynamics of the environment by observing a sequence of actions taken by a teacher. The apprenticeship protocol has been used to efficiently learn flat MDPs and linear MDPs. The KWIK (Li, Littman, and Walsh 2008) or "Knows What It Knows" framework shows that these domains can also be efficiently learned in the autonomous setting. However, there exists a large set of model classes in the Mistake Bound (MB) learning class (Littlestone 1987) that are not efficiently learnable in an autonomous RL setting (e.g. conjunctions of *n* terms). Another critical gap to be addressed in such LbD techniques is the constant need for human intervention and the need for a learning-time guarantee. We have developed a generalized technique in Walsh et al. (2010) that expands the apprenticeship protocol to efficiently learn a wider array of model classes. These classes include all KWIK-learnable classes and all Mistake Bound (MB) classes. We characterize the efficiency of learning by limiting the number of human interactions required.

Exhibited Demonstration

Our demonstration is based on a conjunctive learning task called the Taxi domain, as described in Diuk, Cohen, and Littman (2008). Using a teacher's help, the robot's task is to start from a random state, locate and navigate to an object (passenger), pickup the object and manoeuver it to the destination and drop-off the object at the goal. The object and goal positions are chosen randomly from a set of 3 predefined positions. We begin by describing the different components used and the basic setup of the demonstration.

Our robot, shown in Figure 1 is constructed using a Lego Mindstorms™ kit. It had a standard dual-motor drive train and a third motor controlling its "gripper" that descends over the object and holds it next to the robot's body, and also raises up to "drop-off" the object. The object is a standard 2.2-inch Rubik's Cube initially placed away from the robot. The locations of the robot, the object and the goal are identified and tracked using an overhead camera over a $40in \times 40in$ flat platform. The discrete actions for the robot allow it to move forward, backward, turn right, turn left, and pick-up or drop-off the object (lower or raise the gripper). For the pick-up action, which needs to be quite precise, the atomic action has a subroutine that causes the robot to move forward slightly and then, if its touch sensor is activated, lowers the gripper. Thus, the action is only successful if the agent is facing in the correct direction and near the object, otherwise it just moves forward slightly. The laptop sends these action commands to the robot and receives feedback from the robot via Bluetooth.

Demonstrations are given by a human teacher in the form of a *trace*. A trace is a sequence of states, actions and rewards obtained by executing the teacher's policy from the initial start state to the goal state. These traces are given by the user controlling the robot directly from the laptop. The robot acts autonomously for H steps according to its own policy and, if at any point during the episode the robot performs suboptimally, the user provides a trace at the end of that episode. Therefore, in our protocol, the traces are not given upfront, but are given when required based on the teacher's observation of the robot's policy. This way the learner observes samples of real world transitions *and* re-

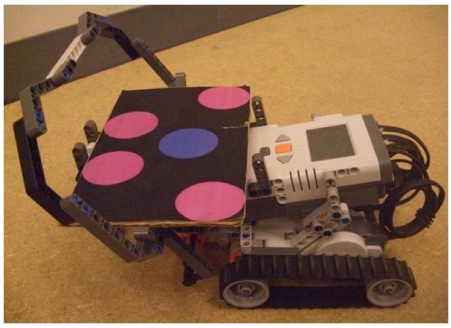

Figure 1: The gripper robot with the colored tracking fiducial.

wards collected by the teacher and uses this "experience" in a traditional model-based RL fashion. In our implemented system, we say that robot has learned the task if it is able to complete the task twice on its own from random start states.

Algorithm Description

In the Taxi world the robot along with the help of a human teacher is required to learn the various conditions of the task and the effects of its actions. Under the generalized apprenticeship learning protocol, we show that this task can be efficiently learned when a "Smart" human interacts with a model-learning robot.

We first describe the state representation that we have used for our task. Object-oriented MDPs (OOMDPs), as described by Diuk, Cohen, and Littman (2008), are used to represent the state space of the environment. OOMDPs are made up of objects with attributes (e.g., their x, y position coordinates) and predicates that must be defined in terms of these attributes (e.g., *On*(A,B): $A.y = B.y + 1$). Actions are described by condition-effect pairs such that in state s_t, the condition (a conjunction over the predicates) that holds (conditions may not overlap) governs which effect occurs. The effects themselves describe changes to the objects' attributes. In our setup, using the OOMDP representation, the attributes of the robot are its x and y coordinates and its orientation θ. The predicates for the OOMDP conditions were *WallToLeft*, *WallToRight*, *WallBehind*, *WallInFront* (all relative to the robot) as well as *HasPassenger*, and *PassengerReachable*, and *AtDestination*. The OOMDP effects were changes in the robot's attributes and the status of the passenger. The rewards of the domain were set as -1 per step and 0 for a successful dropoff at the correct location.

The dynamics of such an environment are learned using a model-learning framework that we call the Mistake-Bounded Predictor (MBP) agent. The robot learner takes actions in the world and learns the effects. Using those effects, the learner builds a model of the world by making predictions. At a high level, the MBP-agent learns in manner similar to the agent described by Diuk, Cohen, and Littman (2008) except that it does not actively explore and uses a Mistake Bound model to learn the transitions. The MBP-Agent, when asked to fill in transitions for conditions it has not witnessed, predicts a "no change" transition (a pessimistic outcome). This way the MBP learner never acknowledges uncertainty, it believes whatever its model tells it (which could be mistaken). While autonomous learners run the risk of failing to explore under such conditions, the MBP-agent can instead rely on its teacher to provide experience in more "helpful" parts of the state space, since its goal is simply to do at least as well as the teacher. The MBP robot learns a model of the world by making only a polynomial number of mistaken predictions (Walsh et al. 2010).

The teacher helping the MBP robot learn a model of the world, is a "Smart" human. In our protocol, a Smart human is one who provides a *Valid Trace*. A Valid trace is defined as the one in which the human teacher does better than what the agent thought was best. It teaches the robot something new about its world and helps the robot reach the desired goal. Introducing a teacher into the learning loop in this manner allows us to characterize the learning efficiency as *PAC-MDP Trace*. This provides a polynomial bound on the number of smart-human interactions required to guarantee efficient learning (Walsh et al. 2010).

Conclusions

This report describes the implementation of our generalized apprenticeship learning protocol in a real world environment. In our demonstrations, the robot was able to learn the taxi task using only a single demonstration from the human teacher. We were able to guarantee efficient learning with limited smart-human interactions and limited mistakes by the robot. The challenges we faced were primarily related to overcoming real world stochasticity and the ability to efficiently learn the conditions given the set of possible effects for each action. As a part of future work, we would like to extend the protocol to other learnability classes (combinations of KWIK and Mistake Bound) and apply it to a number of real world domains. We are currently pursuing work on apprenticeship learning in robotics and natural language processing.

Acknowledgments

We would like to thank Thomas Walsh and Carlos Diuk for their contributions towards building the foundations of this work. We would also like to thank Chris Mansley for providing the software used for object tracking.

References

Abbeel, P., and Ng, A. Y. 2005. Exploration and apprenticeship learning in reinforcement learning. In *ICML*.

Diuk, C.; Cohen, A.; and Littman, M. L. 2008. An object-oriented representation for efficient reinforcement learning. In *ICML*.

Li, L.; Littman, M. L.; and Walsh, T. J. 2008. Knows what it knows: A framework for self-aware learning. In *ICML*.

Littlestone, N. 1987. Learning quickly when irrelevant attributes abound:A new linear-threshold algorithm. *Machine Learning* 2(4):285–318.

Walsh, T.; Subramanian, K.; Littman, M.; and Diuk, C. 2010. Generalizing apprenticeship learning across hypothesis classes. In *ICML*.