Apprenticeship Learning About Multiple Intentions

Monica Babes babes@cs.rutgers.edu
Vukosi Marivate vukosi@cs.rutgers.edu
Michael Littman mlittman@cs.rutgers.edu
Kaushik Subramanian kau.subbu@gmail.com

September 24, 2010

Abstract

In this paper, we apply tools from inverse reinforcement learning (IRL) to the problem of learning from (unlabeled) demonstration trajectories of behavior generated by varying "intentions" or objectives. We derive an EM approach that clusters observed trajectories by inferring the objectives for each cluster using any of several possible IRL methods, and then uses the constructed clusters to quickly identify the intent of a new trajectory. We show that a natural approach to IRL—a gradient ascent method that modifies reward parameters to maximize the likelihood of the observed trajectories—is successful at quickly identifying unknown objectives. We demonstrate these ideas in the context of apprenticeship learning by acquiring the preferences of a human driver in a simple highway car simulator.

1 Introduction

Apprenticeship Learning [1] (AL) addresses the task of learning a policy from expert demonstrations. In one well studied formulation, the expert is assumed to be acting to maximize a reward function, sometimes assumed to be a linear combination of a set of known features, but the reward function is unknown to the apprentice. The only information available concerning the expert's intent is the set of trajectories from the expert's iteration with the environment and, sometimes, the set of features that make up the expert's reward function. From this information, the apprentice strives to derive a policy that performs well with respect to this unknown reward function. If the apprentice's goal is to also learn an explicit representation of the expert's reward function, the problem is called inverse reinforcement learning (IRL) or inverse optimal control. IRL and AL algorithms take as input a Markov decision process (MDP) without a reward function, and the observed expert behavior, in the form of sequences of state-action pairs, assumed to be generated by an optimal policy in the MDP with respect to the unknown expert reward function. The goal in AL is to find a policy that performs well with respect to the expert's reward function. The goal in IRL is to find a proxy for the expert's reward function. When operating under the assumption that this reward function is a linear combination of a known set of features, IRL translates into the problem of finding the weights of these features that approximate the true weights as closely as possible.

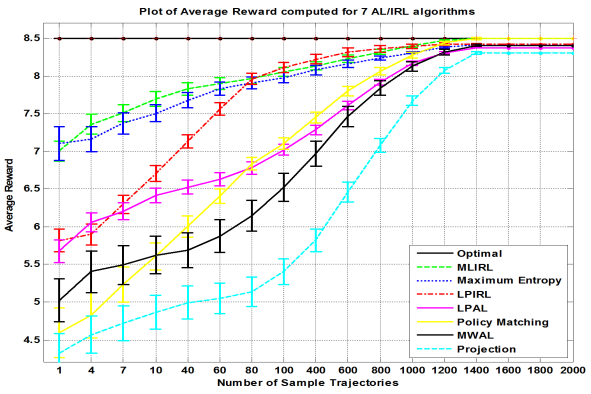

Figure 1: A plot of the average reward computed with increasing number of sample trajectories.

2 MLIRL

We propose a simple IRL method aiming to find the reward function that maximizes the likelihood of the observed trajectories. We call this algorithm Maximum Likelihood Inverse Reinforcement Learning (MLIRL). MLIRL uses the likelihood of the observed expert trajectories, and iteratively changes its parameters (the feature weights) in the direction of the gradient of this likelihood. Therefore, each iteration changes the feature weights to increase the likelihood of the observed behavior. One of the main advantages of MLIRL over existing approaches is that it can provide an acceptable reward function from a very limited number of trajectories, whereas other IRL algorithms require a relatively large number of trajectories as input, in order to be able to provide a good approximation of the unknown reward function (Fig. 1).

3 Learning about Multiple Intentions

In the problem of apprenticeship learning about multiple intentions, we assume there exists a finite set of K or fewer intentions each represented by reward weights θ_k. The apprentice is provided with a set of N trajectories $D = \{\xi_1, ..., \xi_N\}$. Each trajectory is generated by an expert with one of the intentions. An additional trajectory ξ_E is the test trajectory—the apprentice's objective is to produce behavior π_A that obtains high reward with respect to θ_E, the reward weights that generated ξ_E. Many possible clustering algorithms could be applied to attack this problem. We show that Expectation-Maximization (EM) [2] is a viable approach as a streightforward approach to computing a maximum likelihood model in a probabilistic setting in the face of missing data. The missing data in our case are the cluster labels—the mapping from trajectories to one of the intentions.

4 Future Work

Having shown that an EM clustering approach can successfully infer individual intentions from a collection of unlabeled trajectories, we next intend to pursue using these learned intentions to predict the behavior of and better interact with other agents in multiagent environments.

References

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2004.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.