Leveraging Anytime Classifiers to solve POMDPs

Erick Chastain, Rajesh Rao

November 7, 2006

Abstract

In the spirit of RL-Classifier reductions [1], we present a method to reduce a specific class of POMDPs to a time-sensitive family of classifiers. POMDPs are in general intractable to solve exactly, so this reduction allows us to solve the full POMDP by using algorithms like Q-learning, which are known to have polynomial convergence time. A great advantage of this approach is that we can exploit classifiers to discretize and approximate the otherwise continuous belief state. The family of classifiers we introduce are so-called anytime classifiers, which guarantee increasing accuracy (decreasing error rate) when given more training data. In real-time reinforcement learning, speed is important, so it is necessary to determine the optimal time to stop training the anytime classifier. This is called the optimal stopping time. We present an algorithm to learn the optimal stopping time for a given anytime classifier and show its connections to neural mechanisms of sensory decision-making. The type of POMDP which can be reduced to an anytime classifier has states which correspond directly to actions augmented by two more actions: continuing and stopping training of the state classifier. The parallels to optimal stopping problems are clear, and we use an extension of Tsitsiklis & Van Roy's work [2] to derive a learning algorithm for stopping time similar to Q learning. Because all classifiers which are based on calculating the posterior probability using Bayesian models are anytime classifiers (which we show using the Laplace Method), solving this class of POMDP can also be viewed as Bayesian Inference with a stopping rule, with the posterior distribution being parametrized by stopping time.

References

[1] J. Langford and B. Zadrozny. Reducing T-step reinforcement learning to classification. *Proc. of the Machine Learning Reductions Workshop*, 2005.

[2] JN Tsitsiklis and B. van Roy. Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives. *Automatic Control, IEEE Transactions on*, 44(10):1840–1851, 1999.