Preference-based reinforcement learning: a formal framework and a policy iteration algorithm

Johannes Fürnkranz · Eyke Hüllermeier · Weiwei Cheng · Sang-Hyeun Park

Received: 7 November 2011 / Revised: 28 June 2012 / Accepted: 6 July 2012 / Published online: 10 August 2012 © The Author(s) 2012

Abstract This paper makes a first step toward the integration of two subfields of machine learning, namely preference learning and reinforcement learning (RL). An important motivation for a preference-based approach to reinforcement learning is the observation that in many real-world domains, numerical feedback signals are not readily available, or are defined arbitrarily in order to satisfy the needs of conventional RL algorithms. Instead, we propose an alternative framework for reinforcement learning, in which qualitative reward signals can be directly used by the learner. The framework may be viewed as a generalization of the conventional RL framework in which only a partial order between policies is required instead of the total order induced by their respective expected long-term reward.

Therefore, building on novel methods for preference learning, our general goal is to equip the RL agent with qualitative policy models, such as ranking functions that allow for sorting its available actions from most to least promising, as well as algorithms for learning such models from qualitative feedback. As a proof of concept, we realize a first simple instantiation of this framework that defines preferences based on utilities observed for trajectories. To that end, we build on an existing method for approximate policy iteration based on roll-outs. While this approach is based on the use of classification methods for generalization and policy learning, we make use of a specific type of preference learning method called