Learning from Conflicting MDP and Human Reinforcements

Kristof Van Moffaert
Vrije Universiteit Brussel
Pleinlaan 2
1050 Brussels, Belgium
kvmoffae@vub.ac.be

Yann-Michaël De Hauwere
Vrije Universiteit Brussel
Pleinlaan 2
1050 Brussels, Belgium
ydehauwe@vub.ac.be

Peter Vrancx
Vrije Universiteit Brussel
Pleinlaan 2
1050 Brussels, Belgium
pvrancx@vub.ac.be

Ann Nowé
Vrije Universiteit Brussel
Pleinlaan 2
1050 Brussels, Belgium
anowe@vub.ac.be

This research is supported by the IWT-SBO project PERsonalized Products Emerging from Tailored User Adapting Logic (PERPETUAL) (grant nr. 110041).

Abstract

Current approaches in reinforcement learning that combine MDP reward with human reinforcements assume both signals to be complementary. They rely on the fact that the system and the human have the same desire of how a particular task should be completed. We present the case where these reward signals are conflicting, i.e. the goals of the agent are not aligned with the interests of the human trainer. More precisely, we describe an approach how such problems can be solved by multi-objective reinforcement learning.

Introduction

Over the years, learning from humans has received a great deal of attention. In the TAMER framework [1], a reinforcement learning (RL) agent learns a combination of MDP reward and human reinforcement. This framework, and many others, assume that the desired behavior of the human trainer and the MDP reward is *complementary*. This means that feedback of the human trainer can be used to guide the agent in learning the optimal policy more quickly than it would in the case of relying only on the MDP reward. The overall goal of the TAMER framework is to speed up the learning process and to reduce the sampling complexity. In this paper we consider the scenario where the human and the system have different interests, which the agent has to leverage.

Conflicting rewards

Let us consider a controller for systems that interact with a human on a daily basis, such as a household heating appliance. What policy should the controller follow to operate *intelligently*? Does 'intelligently' mean minimizing energy consumption or does it signify maximizing the comfort level of the user? Using the first definition, the energy bill will be minimized as the heater will rarely be turned on, while by the second definition the heater will operate almost constantly, even at moments in time when even when large increases in energy usage create negligible increases in comfort. It is clear that either of these definitions are extreme and will most likely not be what the human is looking for. Therefore, the energy consumption and the human interests are two separate, *conflicting* objectives and the human will most likely be interested in trade-off solutions that compromise these objectives. As we are dealing with systems that interact directly with the end-user, it is crucial that the learning process does not impose large discomfort on the user. Therefore, in contrast to TAMER, our setting does not aim to speed-up the learning process. We are interested in learning in collaboration with the end-user, by optimizing a performance criterion (e.g. energy consumption) while also maintaining a desired level of user satisfaction. In Van Moffaert et al. [2], we apply this idea to intelligently control office equipment, by determining appropriate schedules (on and off times). In this application, the device consists of an office espresso machine, which is often left running 24/7, especially in working environments. We learn trade-off solutions in a multi-objective RL setting by combing both MDP reward and human rewards. The MDP reward represents the energy consumption of a particular mode, which can easily be measured by appliance monitors. In a real-life situation, it is unlikely that the human will spend a lot of time and effort in rewarding the agent for satisfying actions. When the system performs 'good' for the human, he usually finds this obvious or evident behavior. Only when the system is not performing as expected, the human will intervene. Therefore, we only consider negative human reinforcements. Whenever the user is not pleased with the outcome of the system, it will let this know by manually overriding the current control policy. For instance, when the machine is turned off and a beverage is requested, the user manually overrules the agent and waits for the water to reheat. This particular intervention is considered negative user feedback, which is to be minimized in future schedules. By conducting experiments with two sets of weights on each of these two objectives, we obtain two distinct trade-off schedules, i.e. a so-called *energy-oriented* and *user-oriented* schedule. We compare them to a naive always-on schedule in Table 1.

Table 1: The economical properties of the three schedules.

	Always-on	User-o	Energy-o
Hours / day	24h	8h	2h50
Cost / year (€)	578.56	192.86	68.25
Manual overrides	0	1.2	2.1

We see that the potential gains in economical cost of both learned schedules are significant compared to the always-on policy. As learning proceeds, we observe that the amount of human interventions decreases. Surprisingly, the number of overrides of the final policy remains quite low for the energy-oriented schedule as well, as the most busy timeslots are covered.

References

[1] Knox, W. B., and Stone, P. Reinforcement learning from simultaneous human and MDP reward. In

Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS) (June 2012).

[2] Van Moffaert, K., De Hauwere, Y.-M., Vrancx, P., and Nowé, A. Reinforcement learning for multi-purpose schedules. In *5th International Conference on Agents and Artificial Intelligence, ICAART* (2013).