Reinforcement Learning with Stochastic Reward Machines

Jan Corazza[1], Ivan Gavran[2], Daniel Neider[3]

([1]) University of Zagreb (former)
([2]) Informal Systems
([3]) Research Center Trustworthy Data Science and Security, TU Dortmund
(Work conducted at Max Planck Institute for Software Systems)

Despite its great success, reinforcement learning struggles when the reward signal is sparse and temporally extended (e.g., in cases where the agent has to perform a complex series of tasks in a specific order). To expedite the learning process in such situations, a particular form of finite-state machines, called reward machines, has recently been shown to help immensely. However, designing a proper reward machine for the task at hand is challenging and remains a tedious and error-prone manual task.

Recent approaches intertwine reinforcement learning and the inference of reward machines, thereby eliminating the need to craft a reward machine by hand. For example, one successful method transforms the inference task into a series of constraint satisfaction problems that can be solved using off-the-shelf SAT solvers. However, reward machines only model deterministic rewards. When the machine is not known upfront, existing learning methods prove counterproductive in the presence of noisy rewards, as there is either no reward machine consistent with the agent's experience, or the learned reward machine explodes in size, overfitting the noise.

In this talk, we will present recent work that is aimed at addressing the issues of reward machine inference under the presence of noisy rewards. We will introduce the notion of stochastic reward machines, together with a novel algorithm for learning them, and discuss several motivating examples. Stochastic reward machines (SRMs) generalize the notion of reward machines and provide a suitable target for inference algorithms in noisy settings. Our SRM inference algorithm is an extension of the aforementioned constraint-based formulation, and further enhances explainability by recovering information about reward distributions together with the finite-state structure.

We will also briefly discuss the theoretical properties of our algorithm, and present experimental results demonstrating that our algorithm outperforms both a proposed baseline method and existing alternatives on noisy environments, while not worsening performance in the case of deterministic rewards.