Reward Machines for Reinforcement Learning based Gantry Robot Scheduling

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Masterarbeit von Patricia Christin Coberger, Referent: Prof. Dr. Horst Zisgen, Korreferent: Prof. Dr. Frank Bühler

Motivation & Research Question

Since reinforcement learning agents have to be retrained because of a changing environment, it is important to do this fast. Hence, a way was discovered to accelerate the learning process. Reward machines describe a way to achieve the aim of acceleration. However, Icarte et al. [1] who developed the concept of reward machines demonstrated its capabilities mainly on episodic tasks. Therefore, this master thesis aims to find an optimal policy for a continuous reinforcement learning problem using reward machines faster than standard learning approaches without reward machines. On the one hand, the suitability of the general reward machine concept to continuous tasks is investigated. On the other hand, the applicability of algorithms developed for the use of reward machines to continuous problems is analysed.

Applicability of Reward Machines for Continuous Tasks

The general definition of reward machines is not applicable to continuous tasks because infinite input sequences cannot reach an accepting state. Therefore, an accepting condition must be applied that works for infinite inputs. In addition, the CRM-algorithm enforces an early termination as soon as a terminal state is reached. To prevent myopic learning and to allow greater collection of useful information for continuous tasks, it seems reasonable to bypass early termination.

Adjusted Reward Machine Approach

The set of terminal states of reward machines F has to be extended by a Muller acceptance condition in order to also allow infinite input sequences. The Muller condition accepts an input sequence if the set of infinitely frequent visited states is a set in $\mathcal{F}$. Hence, the acceptance condition must be replaced by a combination of F and $\mathcal{F}$.
Furthermore, the CRM-algorithm should be extended to avoid early terminations. This requires the handling of actions when reaching a terminal state. Hence, the state-transition function must be extended by $\delta_f : F \times 2^P \rightarrow U \backslash F$ to allow transitions from a terminal state back to non-terminating states. However, this modification does not effect the experiences that are added to the replay memory (an experience that reaches a terminal state, but none that transitions from a terminal state to a non-terminating state). Besides, a transition from a terminal state back to a non-terminating state should not impact the agent's behaviour. So, these transitions are rewarded with zero.
A reward machine (RM) for continuous tasks is defined as tuple $R = (U, 2^P, \delta, \delta_r, u_0, Acc)$ with

- $\delta : \begin{cases} \delta_u & \text{if } U_t \in U \setminus F \\ \delta_f & \text{if } U_t \in F \end{cases}$ as state-transition function
- $\delta_r : \begin{cases} U \times 2^P \rightarrow \mathbb{R} & \text{if } U_t \in U \setminus F \\ 0 & \text{if } U_t \in F \end{cases}$ as output function
- $Acc = F \cup \mathcal{F}$ as acceptance condition.

Background - Reward Machines and CRM-Algorithm

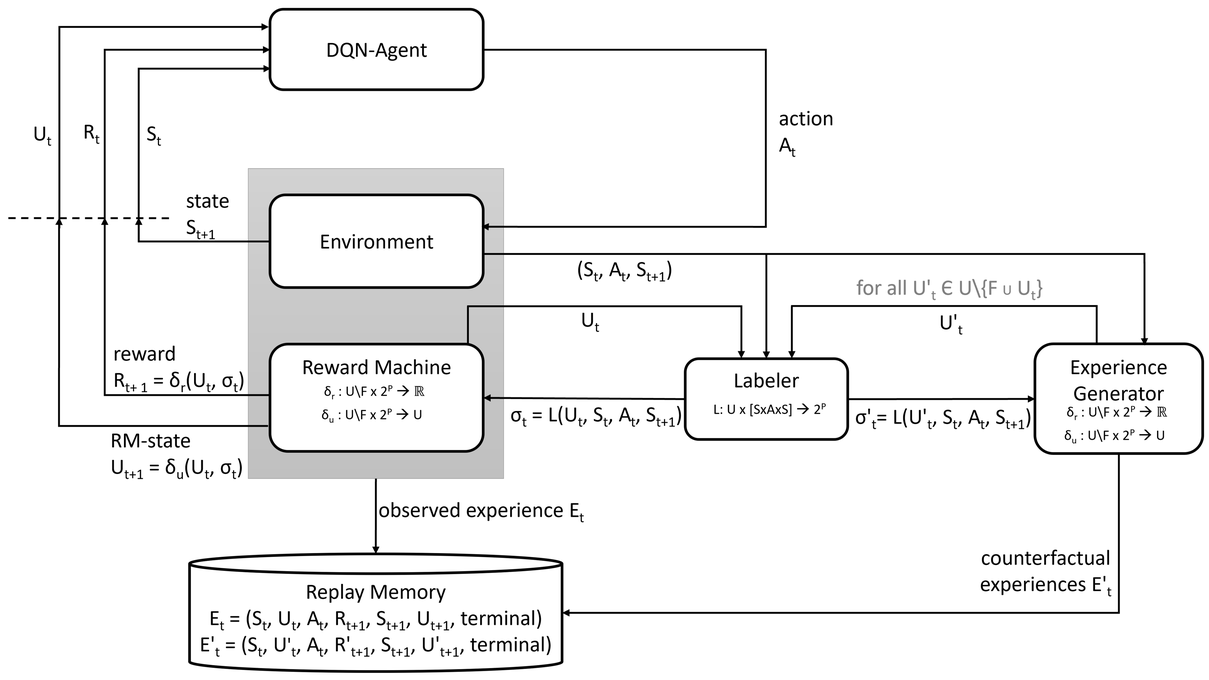

Figure 1:Agent-Environment-Reward Machine interaction with CRM

A reward machine (RM) is defined as tuple
$R = (U, 2^P, \delta_u, \delta_r, u_0, F)$

- U represents the set of states
- 2^P is the input alphabet
- $\delta_u : U \setminus F \times 2^P \rightarrow U$ represents the state-transition function
- $\delta_r : U \setminus F \times 2^P \rightarrow \mathbb{R}$ is the output function
- u_0 defines the initial state
- F represents the accepting states

A reward machine is a finite state machine that reveals the structure of the reward function. The states of the reward machine compress states of the environment due to high-level events in its history or due to reward-relevant aspects in the environment. Logical formulas serve as input to the reward machine and thus, enable a state transition.
The knowledge about the reward function is used by means of a counterfactual reasoning approach (CRM-algorithm) as illustrated in Figure 1. This approach generates an experience for each non-terminating reward machine state that was not observed in the current iteration. Thus, when an action is performed, it can be synthetically observed how this action affects different states of the reward machine. However, as soon as a terminal state of the reward machine is reached, the agent-environment interaction is cancelled and the current episode terminates early.

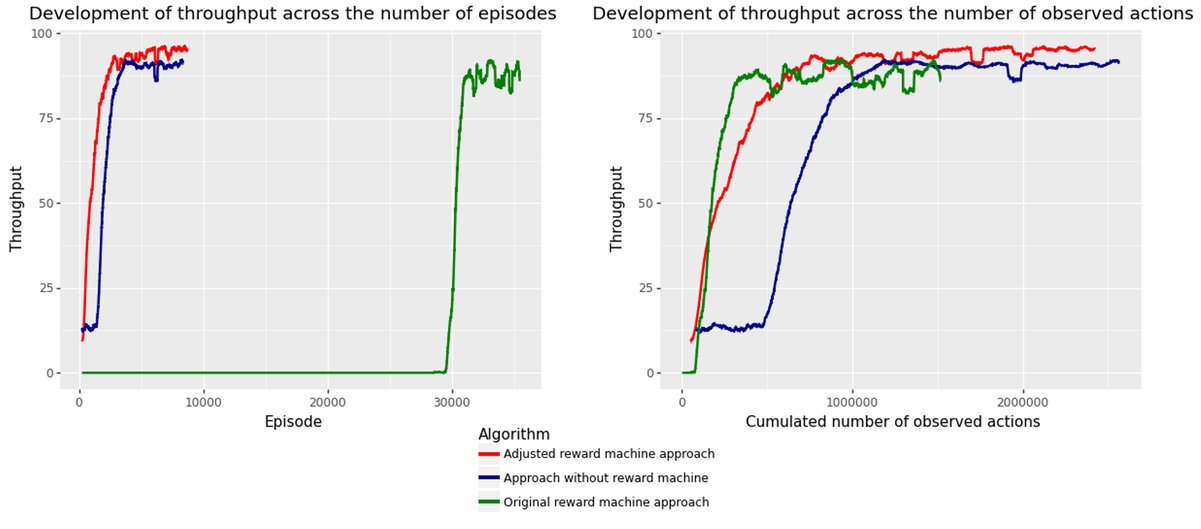

Figure 2:Comparison based on the number of episodes and the number of executed actions

Results

Both the original and the adjusted reward machine approach were able to learn a policy that solves a task for gantry robot scheduling. However, the initial replay memory filled by the original approach had less variety. Also, due to the early termination, only few actions could be executed per episode. Although both approaches were able to solve the task, differences in learning stochastically influenced states became apparent. Especially, the original CRM-algorithm had difficulties learning the optimal behaviour when such a state occurred. Figure 2 illustrates that the adjusted reward machine approach was able to outperform the original reward machine approach and the approach without reward machines in terms of the number of episodes. Furthermore, both reward machine approaches were able to learn an optimal policy with fewer observed actions than the approach without reward machine. This can be attributed to the additional counterfactual experiences that were not observed but were collected.

Conclusion

An analysis of the reward machine definition revealed that it is not applicable for continuous tasks due to the acceptance condition. Furthermore, the CRM-algorithm (and as a consequence the reward machine definition) was extended for use in continuous tasks. Experiments revealed that the original CRM-algorithm would work even without the proposed modifications. Nevertheless, the modifications in the CRM-algorithm support its application for continuous reinforcement learning problems, as fewer episodes are needed for training and the early occurrence of stochastic aspects can be facilitated. Moreover, the adjusted reward machine approach learns a policy faster in terms of the number of episodes (than the original reward machine approach) and the number of observed experiences (than the approach without reward machine).

References

[1] Rodrigo Toro Icarte, Toryn Q Klassen, Richard Valenzano, and Sheila A McIlraith.
Reward machines: Exploiting reward function structure in reinforcement learning.
Journal of Artificial Intelligence Research, 73:173–208, 2022.

[2] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al.
Human-level control through deep reinforcement learning.
nature, 518(7540):529–533, 2015.

[3] Richard S Sutton and Andrew G Barto.
Reinforcement learning: An introduction.
MIT press, 2018.

This Master Thesis was conducted in the context of the research project KiSPo.

LOEWE
Exzellente Forschung für Hessens Zukunft
HESSEN
Hessisches Ministerium für Wissenschaft und Kunst

This project (HA project no. 1286/21-187) is financed with funds of LOEWE – Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz, Förderlinie 3: KMU-Verbundvorhaben (State Offensive for the Development of Scientific and Economic Excellence).