Reward Signal: Discussion of Modeling Human Utility

Chenhao Zhou
Yuanpei College
Peking University
zhouch@stu.pku.edu.cn

Abstract

Utility theory, originally served as a definition for modeling human decision-making process, has long been considered an internal estimate of human choices. The reward signal as an effective assessment of human utility is widely applied in reinforcement learning. In this essay, we will divide the construction of reward signals into two types to discuss: explicit and implicit. Additionally, primarily in the context of reward signals, there are surely a number of debate about whether it fully represents the human utility.

1 Introduction

Utility is generally used to model worth or value. Experiments on infants found that young children can infer and learn others' utility as a core social cognition capability [8]. For computational framework in AI, a basic assumption is that rational agents make decisions to maximize their own utility, while the long-term accumulation of reward signals is served as an estimate of human utility system. Reward plays a significant role as a general purpose signal: Silver et al. [9] assumed that for any desired behavior, task, or other characteristic of agency, there must exist a reward signal that can incentivize an agent to learn to realize these desires.

The utility maximization theory is taken as a backdrop assumption when solving decision-making problems through reinforcement learning [1]. It states that all of what we mean by goals and purposes can be well thought of as maximization of the expected value of the cumulative sum of a received scalar signal (reward). Thus the utility function yields a preference, or rather, partial order, over a set of decisions, and rational agents are expected to act according to such a preference.

Practically we can define an explicit reward function in advance or ask agents to learn an implicit one from human perspectives. The explicit construction of reward function will be discussed in Sec. 2. What followed is the overview of implicit reward learning in Sec. 3. Finally in Sec. 4 we will review some debates on the perspective of reward signal.

2 Explicitly define reward functions

Manual designation of reward function is challenging and closely related to the efficiency of reinforcement learning. So this section we focus on the operations on explicit reward functions.

For the specifying complex task with multiple objectives and safety constraints, explicitly, Jothimurugan et al. [7] propose a language for specifying complex control tasks, along with an algorithm that compiles specifications in the language into a reward function and automatically performs reward shaping.

In order to further utilize the advantages of explicit reward function, Icarte et al. [5] show the reward function's code to the reinforcement learning agent so it can exploit the function's internal structure to learn optimal policies in a more sample efficient manner. They propose reward machines, a type of finite state machine that supports the specification of reward functions while exposing reward function structure.

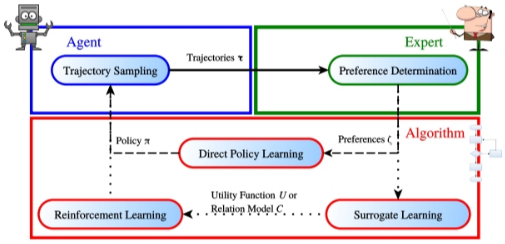

Figure 1: Learning policies from preferences via direct (dashed path) and surrogate-based (dotted path) approaches.

The prior knowledge is put into the definition of the reward function. However, utility represented in hand-crafted functions results in high sensitivity and instability of the agent's learned actions, as well as a possible deviation from the tasks that humans originally expect the agent to perform. Finally it would cause the misalignment between human values and the objectives of the reinforcement learning systems. So we shift to the the implicit reward functions next section.

3 Implicitly learn reward functions

It's often the case that the utility function is dynamic and involves lots of hidden factors [6]. Thus it is usually difficult to hand-specify what correct reward function is for a task. Learning an implicit reward function from different forms of human behaviors, which is the classical problem of inverse reinforcement learning, could be an efficient way of modeling human utility.

3.1 Example-based reinforcement learning

Considering the difficulty of manually design a reward function, and trying to make task specification in reinforcement learning data-driven, Eysenbach et al. [3] derive a control algorithm that directly learns a value function from transitions and successful outcomes, without learning this intermediate reward function. They design a recursive classification to estimate the probability of reaching a success example in the future and optimizes a policy to maximize this probability of success. As for the data-driven learning, the process is end-to-end, may better captures the essence of many real-world control problems.

3.2 Preference-based reinforcement learning

In place of explicit rewards, different kinds of expert's preferences can be used to evaluate an agent's behaviors. Thus preference-based reinforcement learning (PbRL) was proposed, in which an agent receives evaluative signals in the form of preferences over states, actions, or trajectories by interacting with a Controlled Markov Process (CMP) [10].

The typical process of PbRL is composed of several components, as illustrated in Fig. 1. While Christiano et al. [2] obtain the goals defined in terms of human preferences between pairs of trajectory segments. They collect human's comparison among possible trajectories of the agent, thus using that data to learn a reward function through deep reinforcement learning method.

The PbRL is a suitable tool for learning from qualitative, non-numeric rewards, it can provide solutions in domains where numeric feedback is not readily available. However, current approaches require a high amount of preferences or well defined feature spaces, either on a policy or trajectory level. And algorithm may fail in some cases of incomparabilities.

3.3 Instruction-grounding reinforcement learning

In order to build learning algorithms that will tell agents what we want them to do, Fu et al. [4] propose language-conditioned reward learning (IC-RL), which grounds language commands as a reward function represented by a deep neural network. Experiments have shown that conversion from language-defined goals into reward functions can obtain higher ability to generalize to new environments than simply mapping language command to policies.

In this manner, the reward functions are generated based on human-designed language instructions. Thus the agent can learn how to plan and perform the task on its own via reinforcement learning. However the method restricts the training to tractable domains with known dynamics.

Another inverse reinforcement learning or imitation learning approaches which learn from demonstrations, are not directly applicable to behaviors that are difficult for humans to demonstrate (such as controlling a robot with many degrees of freedom but non-human morphology).

4 Discussion about reward signal

The process of reward maximisation is tightly bound to the environment, as diverse reward signals are obtained from complex environments [9]. Yet it is extremely difficult to construct a generic form of reward signal. Rather than maximising a generic objective defined by cumulative reward, practically the signal is often formulated separately for separately for different cases. Besides, the reward signal inevitably loses some exact depict cue from extrinsic and intrinsic contexts.

As an internal attribution, human utility models the expected value derived from people's minds. A general value system should consider prior knowledge, human feedback, interactive use, intrinsic motivation, together forming a hierarchical value function, which may be a more appropriate way.

5 Conclusion

In this essay, we divide the construction of reward signals into two types to discuss. And both the explicit and implicit forms inevitably have deficiencies based on the utility maximization theory. The hierarchical modeling of human utility involving cognition information may be an approximate way. Further research is needed to understand better how humans represent and make decisions with utility, towards a unified utility representation.

References

[1] David Abel, Will Dabney, Anna Harutyunyan, Mark K. Ho, Michael L. Littman, Doina Precup, and Satinder Singh. On the expressivity of markov reward. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

[2] Paul F Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[3] Benjamin Eysenbach, Sergey Levine, and Ruslan Salakhutdinov. Replacing rewards with examples: Example-based policy search via recursive classification. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[4] Justin Fu, Anoop Korattikara, Sergey Levine, and Sergio Guadarrama. From language to goals: Inverse reinforcement learning for vision-based instruction following. *arXiv preprint arXiv:1902.07742*, 2019. 2

[5] 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*, 2022. 1

[6] Hong Jun Jeon, Smitha Milli, and Anca Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[7] Kishor Jothimurugan, Rajeev Alur, and Osbert Bastani. A composable specification language for reinforcement learning tasks. *In Advances in Neural Information Processing Systems*, 2020. 1

[8] Shari Liu, Tomer D Ullman, Joshua B Tenebaum, and Elizabeth S Spelke. Ten-month-old infants infer the value of goals from the costs of actions. *Science*, 358(6366):1038–1041, 2017. 1

[9] David Silver, Satinder Singh, Doina Precup, and Richard S. Sutton. Reward is enough. *Artificial Intelligence*, 2021. 1, 3

[10] Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of perference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18: 1–46, 2017. 2