Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)

Optimizing Interactive Systems with Data-Driven Objectives

Ziming Li
University of Amsterdam, Netherlands
z.li@uva.nl

Abstract

Effective optimization is essential for interactive systems to provide a satisfactory user experience. However, it is often challenging to find an objective to optimize for. Generally, such objectives are manually crafted and rarely capture complex user needs in an accurate manner. We propose to infer the objective directly from observed user interactions. These inferences can be made regardless of prior knowledge and across different types of user behavior. It is promising if we model the objectives directly from the user interactions which we use to optimize interactive systems, which will improve user experience and dynamically reacts to user actions.

1 Background and Motivation

An interactive system is a system where a user can interact with a pre-designed system by generating list of history traces. With the emerging demands of intelligent systems, such as dialogue system [Li *et al.*, 2017a] and search engine [Yan *et al.*, 2014], how to design an interactive system becomes more and more popular in computer science. By interacting with the system, the user is getting reward, such as knowledge or information that he wants to gain. To adapt user's needs and tasks, an interactive system must be able to establish a set of assumptions about what kind of interactions or replies will match user's need and satisfy them. The main component of designing such systems is an objective functio, which should reflect the quality of the system and satisfaction of users and is the goal that the system should aim for.

Understanding and modeling user behavior is a fundamental problem for any interactive system as insight into user behavior will lead towards appropriate evaluation: what satisfies user needs and what frustrates users [Li *et al.*, 2017b]. Currently, the common strategy is to use domain knowledge to design objectives, e.g. clicks on search result page [Luo *et al.*, 2015] or the similarity between generated reply and predefined answer in dialogue systems [Li *et al.*, 2016]. However, we know that user behaviour is inherently more complex and should depend on many aspects [Kosinski *et al.*, 2013]. Simply defined objective functions can only reflect part of user's needs and are not able to deal with more complex behaviors as users have different preferences and can display different behavior. To design an appropriate objective function for an interactive system, a strong and comprehensive background about this domain is essential but may still cannot get satisfied performance.

Therefore it sounds appealing to model our objectives directly from the user interactions which we use to optimize the system, which will improve user experience and dynamically reacts to user actions.

2 What We Have Done

We have investigated the following specific questions.

2.1 How to Design Proper Objectives for an Interactive System and Improve System Performance According to the Designed Objectives?

There is a common scenario of how an interactive system works: a user comes with some goal in mind (his reward) and start interacting with the system by making actions and a system puts her in a particular state. Modelling the interactive systems as Markov Decision Process is a possible solution. In the diagram of MDP, reinforcement learning could be a natural choice to find the optimal policy. Here we can regard the user's reward function as the objectives for our interactive systems. We investigated the following subquestions in our paper [Li *et al.*, 2017b; Li *et al.*, 2018]:

1. How to model interactions between users and systems?
2. What structures the user reward functions should have?
3. Which kind of feedback should user interactions supply?
4. How to recover reward functions directly from the user interactions?
5. How to design personalized reward functions for interactive systems?

2.2 Can We Improve User Experience in a Simulated Interactive System Based on Designed Objectives?

As we mentioned above, reinforcement learning is a natural choice to find the optimal policy for agents. However, we

model the user as agent and system as environment at the beginning and our goal is not finding the optimal policy for the user rather than the optimal policy for the system, which decide how to transfer user to the next state after the user has taken one action at current state. We need to transfer the job of finding optimal system to how to find the best policy for the system. There should be a mechanism to change the roles of user and interactive system, which means the system could also be modelled as agent while user as environment. We investigated the following subquestions in our paper [Li *et al.*, 2018]:

1. What is the relation between recovered user reward function and system reward functions?
2. How to model the transition distribution while user is the environment?
3. Which reinforcement learning method is the best choice in terms of performance and efficiency?
4. How to evaluate the difference between the original system and optimized system?

2.3 Can We Optimize a Dialogue System Without Designing Reward Function?

After the last two works, we applied the proposed framework to train a dialogue system which can be regarded as a typical interactive system. However, chat-bot has its own characters and it is not so straightforward to model chat-bot as the setting we proposed above. Besides, chat-bot is much more complex since the state space and action space are huge which will result in one serious problem: how to find the best policy in this situation? The ideal solution should be recovering the reward function first and then applying policy gradient method to search for the optimal policy. This is the normal procedure of inverse reinforcement learning. However, it is quite difficult to learn the reward function directly and in some scenarios we only want the optimal policy.

In our paper [Li *et al.*, 2019], we first applied adversarial imitation learning to search for the optimal policy directly without recovering the reward function. A discriminator is trained to distinguish the machine-generated dialogues from the real human-generated dialogues and provide the policy model with reward signals. However, the reward signal from a poor discriminator can be very sparse and unstable, which may lead the generator to fall into a local optimum or to produce nonsense replies. Then, in the framework of adversarial inverse reinforcement learning, we propose a new reward model for dialogue generation that can provide a more accurate and precise reward signal for generator training. We evaluate the performance of the resulting model with automatic metrics and human evaluations in two annotation settings.

3 What We Are Going to Do

We are going to investigate the following questions.

3.1 How to Explore the Abilities of the Recovered Reward Function, Such As Explainability?

As we said in last subsection, ideally we want to recover the reward function which can explain the demonstrated dialogues and then use this reward function to optimize the system policy. There are several subquestions we are interested:

1. How to measure the quality and reliability of the recovered reward function
2. Can we utilize the recovered reward function to evaluate the performance of other interactive systems from the same domain?
3. Can we utilize the recovered reward function to detect and locate the inappropriate actions taken by the system to analyze and adjuct the system policy.

3.2 How to Optimize a Personalized Interactive System?

In the current setting, this question can be transferred to another question "how to recover personalized reward functions for different users". In the same interactive system, users may have different preferences and this characteristic will lead to very diverse user behaviors.

1. How to incorporate user profile features to a reward function? This will be the first question we need to answer.
2. How to infer users' reward functions with few demonstrated behaviors? This is another challenge we need to deal with.

References

[Kosinski *et al.*, 2013] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *PNAS*, 110(15):5802–5805, 2013.

[Li *et al.*, 2016] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In *EMNLP*, pages 1192–1202, 2016.

[Li *et al.*, 2017a] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. In *EMNLP*, pages 2157–2169, 2017.

[Li *et al.*, 2017b] Ziming Li, Julia Kiseleva, Maarten de Rijke, and Artem Grotov. Towards learning reward functions from user interactions. In *ICTIR*, pages 289–292, 2017.

[Li *et al.*, 2018] Ziming Li, Artem Grotov, Julia Kiseleva, Maarten de Rijke, and Harrie Oosterhuis. Optimizing interactive systems with data-driven objectives. *arXiv preprint arXiv:1802.06306*, 2018.

[Li *et al.*, 2019] Ziming Li, Julia Kiseleva, and Maarten de Rijke. Dialogue generation: From imitation learning to inverse reinforcement learning. In *AAAI*, 2019.

[Luo *et al.*, 2015] Jiyun Luo, Xuchu Dong, and Hui Yang. Session search by direct policy learning. In *ICTIR*, pages 261–270, 2015.

[Yan *et al.*, 2014] Jinyun Yan, Wei Chu, and Ryen W White. Cohort modeling for enhanced personalized search. In *SIGIR*, pages 505–514, 2014.