Possible ways to learn and represent human utility

Yao Yuanhang*

Peking University
2000017813@stu.pku.edu.cn

Abstract

As described in the problem statement, the human utility function is internal to humans, lacks meaningful units of measurement, and varies across individuals. Estimating human utility for different tasks using a general computational framework remains challenging. In the following article, I will explore the characteristics of Preference-Based Reinforcement Learning, Expected Utility Theory, and conventional computational models such as decision trees and Bayesian networks, along with specific methods for learning and representing human utility, and their advantages and disadvantages in data collection, generalization, and efficiency.

1 Introduction

Starting from the definition of human utility, it refers to the ability of goods or services to satisfy people's desires or the degree of satisfaction consumers feel when consuming goods or services. The utility function is a function that maps results or states to the basic value that expresses the agent's needs. In this representation, the utility function assigns a numerical value to each result or state. Human utility is a subjective concept, and different people may assign different utility values to the same task. This difference may be due to individual values, beliefs, cultural backgrounds, experiences, and other factors, so the problem will say that it is difficult to estimate human utility for different tasks using a general computational framework. Therefore, estimating human utility needs to take these factors into account and use methods specific to individuals for estimation.

By thinking about the difficulties of estimating human utility, we can infer what characteristics our proposed estimation algorithm should have. Obviously, the estimation method we propose should have the following characteristics: personalization: the estimation method should consider individual differences and use methods specific to individuals for estimation. Accuracy: the estimation method should estimate human utility as accurately as possible. Interpretability: the estimation method should be able to explain the meaning of the estimated results and the principles behind them. Reliability: the estimation method should be reliable and able to obtain consistent results in different environments. Understanding these requirements makes it easier for us to find the most suitable algorithm framework in subsequent comparisons and experiments.

2 Preference-Based Reinforcement Learning

2.1 Algorithm introduction

As described in the problem statement, preference-based reinforcement learning (PbRL) is a form of reinforcement learning in which the agent's goal is to learn its preferences rather than learning the optimal policy. In PbRL, the agent's goal is to learn a preference function that maps states to preference values. The preference function can be learned by observing the agent's behavior without modeling the agent's internal state. PbRL uses preferences instead of numerical reward

*Use footnote for providing further information about author (webpage, co-first authors, *etc*.).

values in traditional reinforcement learning to better elicit human opinions about the goal, especially when numerical reward values are difficult to design or interpret. Relevant research [3] proposed a finite-time analysis of a general PbRL problem. The researchers first proved that if the trajectory preference of PbRL is deterministic, there may not exist a unique optimal policy. If the preference is stochastic and the preference probability is related to the hidden reward value, they proposed a PbRL algorithm that can identify the optimal policy with high probability through a simulator. The specific method is to explore the state space by navigating to undeveloped states and use a combination of dueling bandits and policy search to solve PbRL.

In addition, PbRL also has the ability of active learning [1]. It can learn from a small number of queries and generalize to new tasks. This method learns the reward function by asking the user's preference instead of observing the user's behavior. During the learning process, the method uses an active learning strategy to select the most useful queries so as to learn the optimal reward function in as few queries as possible. Its implementation method is to learn the reward function from a continuous hypothesis space by maximizing the volume of the hypothesis space removed by each query. The hypothesis space is weighted in the form of a log-concave distribution, and a limit on the number of iterations required for convergence is provided.

2.2 Pros and cons analysis

Based on the detailed description of the preference-based reinforcement learning algorithm above, I can further analyze its pros and cons. Because PbRL emphasizes learning from a small number of queries and generalizing to new tasks, it does not require a large amount of data samples for initial training, and its advantages are also reflected in data collection. It has low requirements for data size, and people only need to collect a small part of the data to provide the model with good estimation results through active learning and reward replacement. At the same time, we can see that PbRL emphasizes actively acquiring knowledge features in a specific aspect, and the subsequent estimation results are also based on these specific preferences, which obviously lacks generalization ability. This is the disadvantage of PbRL. Another disadvantage is that because the preference function is learned based on the agent's feedback, PbRL may be limited by the agent's behavior, because the agent's behavior may not necessarily reflect its true preference. In terms of efficiency, combined with the finite-time analysis of a general PbRL problem proposed by researchers, we can know that it can identify the optimal policy with high probability and has high efficiency.

3 Expected utility theory

3.1 Algorithm introduction

Expected utility theory is a decision theory that views decision problems as choosing actions with the maximum expected utility. The main idea of expected utility theory is that people consider uncertainty factors when making decisions and calculate expected utility based on the probability and utility values of uncertainty factors. Expected utility theory can be used to estimate human utility, but it needs to take into account individual differences and use methods specific to individuals for estimation. Most articles on expected utility theory explore the relationship between Bayesian psychology and human rationality, focusing on the impact of Bayesian psychology on human rationality and the adaptability between Bayesian cognitive models and actual human performance. Researchers believe that Bayesian psychology provides a new way to understand human decision-making behavior[2] and can be used to explain some seemingly irrational aspects of human decision-making behavior.

3.2 Pros and cons analysis

The advantage of expected utility theory is that it can use mathematical formulas to represent the utility function, which does not require a particularly complex representation method, bringing great efficiency. However, the disadvantage is that it needs to model the agent's preferences, which may require a large amount of data (this is different from the modeling of preferences in PbRL, which requires modeling the agent's internal state rather than directly learning behavioral patterns). In addition, expected utility theory may be limited by the agent's risk attitude, because the agent's risk attitude may not necessarily conform to the assumptions of expected utility theory. At the same time,

humans cannot always maintain rationality and pursue maximum expected efficiency, which also brings great trouble to the estimation of the model.

4 Some other estimation methods

4.1 Algorithm introduction

Earlier, we specifically introduced two relatively novel algorithms in the field of human utility estimation, and some more classic and effective estimation methods can also be applied to this field. For example, modern decision theory, decision trees, Bayesian networks, Markov decision processes, and other methods can be used to represent human utility. This section briefly introduces decision trees and Bayesian networks. Decision tree is a non-parametric supervised learning algorithm used for classification and regression tasks. It is a hierarchical tree structure composed of root nodes, branches, internal nodes, and leaf nodes. The decision tree starts from the root node, which has no incoming branches. Then, the outgoing branches of the root node provide information to the internal node (also called the decision node). Both types of nodes perform evaluations based on available functions to form homogeneous subsets, which are represented by leaf or terminal nodes. Leaf nodes represent all possible results in the data set. The application of decision trees in the field of human utility estimation includes various reasoning and analysis modes such as performance evaluation and decision analysis. For example, decision trees can be used to infer decision-making on the parameter performance and combat decisions of helicopters. Bayesian network is a probabilistic graphical model used to represent the dependency relationship between variables. It is a directed acyclic graph, where each node represents a random variable, and the edge represents a probability dependency relationship between variables. It can also be used to model and analyze human utility[4].

4.2 Pros and cons analysis

Regarding the advantages of Bayesian networks, they can handle incomplete data, uncertainty, relationships between multiple variables, model reconstruction, model migration, model fusion, model interpretability, and have good generalization. The disadvantage is that it requires a large amount of data for training, high requirements for data collection, and high computational costs because they need to calculate all possible results.

The advantages of decision trees include ease of understanding and interpretation, ability to handle nominal and numeric data simultaneously, suitability for samples with missing attributes, ability to handle unrelated features, fast running speed when testing data sets, and ability to produce feasible and good results for large data sources. The disadvantages include easy overfitting, low stability, calculation results are locally optimal rather than globally optimal, and are affected by sample imbalance.

References

[1] Yang Gao and Anca D Dragan. Active preference-based learning of reward functions. In *Advances in Neural Information Processing Systems*, pages 101–111, 2018. 2

[2] Thomas L Griffiths, Christopher G Lucas, and Joshua J Williams. *Bayesian Psychology and Human Rationality*. Cambridge University Press, 2018. 2

[3] Yichong Xu, Ruosong Wang, Lin F Yang, Aarti Singh, and Artur Dubrawski. Preference-based reinforcement learning with finite-time guarantees. *arXiv preprint arXiv:2006.08910*, 2020. 2

[4] Xiang Zhang, Xiaoyu Zhang, Xiaoxia Liu, Xiang Li, and Jianping Li. . , 55(1):1–7, 2019. 3