Prototype-Based Off-Policy Evaluation

Presenter:
Anton Matsson

BACKGROUND. Importance sampling (IS) is often used to perform off-policy evaluation (OPE)—estimating the value of a target policy π based on data gathered under a (typically unknown) behavior policy μ—but it suffers from high variance when π is significantly different from μ. Standard practices may be insufficient for domain experts to diagnose problems and assess the quality of an IS value estimate.

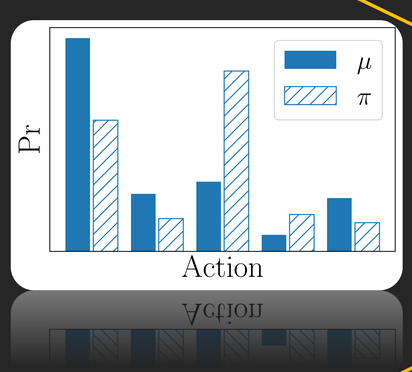
METHOD. We propose estimating the unknown behavior policy using prototype learning [1]. The idea is to express the probability $p_\mu(A|H)$ of taking an action A given an input history H by comparing H to a few examples—prototypes—from the data. The learned prototypes induce a clustering of the data, allowing us to describe inputs for which μ and π differ in *suggested actions* and *resulting values*—providing insights about the IS value estimate.

RESULTS. We demonstrate our method by examining the AI Clinician [2], an AI-based policy for sepsis management. Already at the initial time step we observe a clear difference between π_{AIC} and μ—followed by physicians in data—making it difficult to accurately estimate the value of π_{AIC}. To reduce variance, we evaluate the policy of following π_{AIC} in the initial time step and then following μ. By dividing the IS value estimate into prototype-based contributions, we find that no types of patients greatly benefit from being treated according to π_{AIC} at the onset of sepsis.

References

1. Li, O. et al. Deep learning for case-based reason through prototypes: A neural network that explains it predictions. In *AAAI*, 2018.
2. Komorowski, M. et al. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11): 1716–1720, 2018.

We estimate the behavior policy μ for OPE using prototypes, allowing us to describe differences between μ and the target policy π and their estimated values



Take a **picture** to **download** the **full paper.**

Anton Matsson
Fredrik D. Johansson

CHALMERS
UNIVERSITY OF TECHNOLOGY