New Frontier of Interactive and Interpretable Machine Learning

Jiaming Qu

University of North Carolina at Chapel Hill
jiaming@unc.edu

1 Research Progress

In general, my research as a Ph.D. student is at the intersection of interpretable Machine Learning (ML) and Human-Computer Interaction (HCI). My research started from developing interpretable retrieval models for medical literature search [1,2]. The proposed model not only leverages simple-yet-robust decision tree which has performance not worse (or even better) than black-box methods especially when training data is limited, but also resembles physicians' structured relevance judgment process to preserve transparency and interpretability. After that, I became curious about how the theoretical models could be applied in practice, and thus my research interests switched to studying *the effect of interpretability features on decision-making tasks.* At the very beginning, we launched an exploratory user study on Amazon Mechanical Turk with 200 participants who were asked to scrutinize movie plot summaries predicted to satisfy multiple genres [3]. We found that two designed interpretability features improved users' perceptions and performances, which revealed the importance of providing such interactive explainable tools in faceted search systems when facet-values were automatically assigned. However, such crowd-sourced study was only based on quantitative analysis, and little was known about what users went through when they interacted with interpretability features to make decisions. Therefore, in a recent work (under review) [4], we conducted a lab study ($N = 30$) in which users were asked to make relevance judgments on a medical search task. By analyzing qualitative data collected through stimulated recall protocol and semi-structured interviews, we gain an in-depth understanding of users' internal cognitive processes in a relevance decision task, how ML interpretations influence their cognitive processes, and what is missing in current ML interpretability features.

2 Future Plans

For the next step of my dissertation and research, I aim to continue studying interpretable ML and HCI with a focus on assisting decision-making tasks. In particular, I am interested in some detailed topics as listed below.

(1) **Support sense making in decision making.** Based on our recent work [4], we found that *sense making is a crucial part of decision making*, which was often overlooked in current research. Essentially, ML models learn a function f which maps input X to output Y (i.e., $f : X \to Y$). However, current interpretable ML techniques aim to articulate how a model predicts the output based on specific parts of the input, which helps users see the reasoning logic behind a decision, *not* how to make better sense of the input or output. Therefore, I aim to develop tools to support sense making (i.e., better understand the input X and output Y rather than the function f only) and learn the effects on decision-making performance.

(2) **Facilitate decision making under uncertainties.** Our recent work [3,4] also revealed an interesting phenomena: users became confused on boundary classification cases when the interpretation was vague and ambiguous (i.e., no strong evidence). Current interpretations are commonly designed to show why the model is confident, but not why it is unconfident. However, it is necessary to develop interpretable ML approaches that are aware of low prediction confidence and provide appropriate explanations to to facilitate decision making under uncertainties.

(3) **Counter pitfalls of interpretable ML.** The current interpretable ML community has witnessed much more research that claims the benefits of interpretation, while the pitfalls of interpretation are quite under-explored. However, in our previous work [3,4], we found that interpretation did not necessarily improve decision performance and might even result in worse outcomes. Therefore, I aim to survey research to learn in what scenarios interpretation may fail and the reasons behind, based on which design implications could be proposed to counter those pitfalls.

References

1. Jiaming Qu, Jaime Arguello, and Yue Wang. Towards explainable retrieval models for precision medicine literature search. In *Proceedings of the 43rd ACM SIGIR*, 2020.
2. Jiaming Qu, Jaime Arguello, and Yue Wang. A deep analysis of an explainable retrieval model for precision medicine literature search. In *ECIR*, 2021.
3. Jiaming Qu, Jaime Arguello, and Yue Wang. A study of explainability features to scrutinize faceted filtering results. In *Proceedings of the 30th ACM CIKM*, 2021.
4. Anonymous. Anonymous. In *Proceedings of the 45th ACM SIGIR*, 2022, Under Review.