Intelligible Explanations in Intelligent Systems

Sule Anjomshoae and Kary Främling

Department of Computing Science, Umeå University, Sweden
sule.anjomshoae@umu.se, kary.framling@umu.se

Abstract. EXplainable Artificial Intelligence (XAI) is a growing field of interest for researchers and developers to provide trusted and transparent intelligent systems. We conducted a review on explainable agents to gain insight into how explanations are generated, communicated and evaluated. Then, we presented preliminary results of Contextual Importance and Utility (CIU) method for generating human understandable explanations. This paper highlights important findings from the systematic literature review and presents the utility of CIU method in generating and communicating intelligible explanations.

Keywords: XAI · explanations · contextual importance · contextual utility.

1 Introduction

Explanations for the actions and predictions made by intelligent systems play an important role in adoption of these systems for decision support. Previous surveys on explanations mainly draw attention to data-driven algorithms (i.e., interpretable machine learning model). Little attention has been devoted to goal-driven XAI (e.g., agent, robot) which known for their ability to communicate their decisions in an intelligible way. There is a need for identifying trends and future challenges in goal-driven XAI. We conducted a systematic literature review to provide insights into how current works in this domain solve the problem of (i) generating and (ii) communicating intelligible explanations [1].

In generating explanations, most of the studies follow the introspective informative explanations. This explanation type is based on the reasoning process which leads to a decision [4]. Moreover, several studies suggested generating explanation facilities based on practically relevant theoretical concepts [5]. One of these is providing contrastive justifications (i.e., contrasting instance against the instance of interest) complementary to complete explanations (i.e., causes of an individual prediction).

In communicating explanations, most studies selected single modality to communicate explanations (e.g., text-based [6], visuals [7], speech [8]) which restricts the interaction quality. Addition to that limited works presented personalized (e.g., age, expertise, role) and context-aware (e.g., situation awareness) explanations to provide explanatory information that is relevant to the user and its

context. We presented the CIU method's capability in generating and communicating such explanations [2]. The utility of this method is presented in the next section.

2 Contextual Importance and Utility

Contextual importance and utility method, originally proposed by Främling [3], was adopted to generate explanations for linear and non-linear machine learning model predictions [2]. The result and utility of the CIU method in providing intelligible explanations are briefly highlighted below;

(i) Generating explanations:

- *Model-agnostic explanation method:* CIU method can be applied to a range of linear and non-linear black-box models. This increases the generalizability of the explanation method in selection of the learning model.
- *Complete and contrastive explanations:* Creating contrastive explanations and comparing the differences to another instance can often be more useful than the complete explanation alone for a particular prediction. CIU allows explaining why a certain instance is preferable to another one, or why one class (outcome) is more probable than another.

(ii) Communicating explanations:

- *Personalized explanations:* CIU values can be represented with different levels of details and produce explanations that are tailored to the users' specification. Since the concepts and vocabularies that are used for producing explanations are external to the black-box, the vocabularies can be communicated depending on the user they are intended for.
- *Visualization of the explanations:* CIU method enables to change the representation type if it turns out that the currently used modality is not suitable for the user's understanding. This is what humans tend to do when another person does not seem to understand already tested explanation. The variability in representing explanations could improve the interaction quality, particularly in time-sensitive situations.

3 Conclusion

This extended abstract summarizes our efforts towards providing intelligible explanations for intelligent systems. Future work relates to CIU's utility in producing user-aware and context-aware explanations by taking the user's characteristics into account and investigating the feasibility of the explanations in real-world settings.

References

1. Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable Agents and Robots: Results from a Systematic Literature Review. In: Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (2019).
2. Anjomshoae, S., Främling, K., Najjar, A.: Explanations of Black-Box Model Predictions by Contextual Importance and Utility. In: Proceedings of the 1st International Workshop on EXplainable TRansparent Autonomous Agents and Multi-Agent Systems (2019).
3. Främling, K. and Graillot D.: Extracting Explanations from Neural Networks. In: Proceedings of the ICANN (1995).
4. Sheh, R.K.: Different XAI for different HRI. In: AAAI Fall Symposium Series (2017).
5. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. In: Artificial Intelligence (2018).
6. Broekens, J., Harbers, M., Hindriks, K., Van Den Bosch, K., Jonker, C. and Meyer, J.J.: Do you get it? User-evaluated explainable BDI agents. In: German Conference on Multiagent System Technologies. Springer, Berlin, Heidelberg (2010).
7. Quijano-Sanchez, L., Sauer, C., Recio-Garcia, J.A. and Diaz-Agudo, B.: Make it personal: a social explanation system applied to group recommendations. In: Expert Systems with Applications 76 (2017).
8. Lettl B, Schulte A.: Self-explanation capability for cognitive agents on-board of UCAVs to improve cooperation in a manned-unmanned fighter team. In: AIAA Infotech@ Aerospace (I@ A) Conference (2013).