Master 2 Internship in Artificial Intelligence (IRIT et 3IA ANITI, Toulouse, France)

«Explainability of Preference Models for decision-aid system »

Decision aid systems, like on-line configurators or recommender systems, need to adapt themselves to each user in order to offer a better interaction and guide the user quickly to the best decision for her: the system should be able to gather a model of the preferences of the user, and be able to show her, almost instantly, during the interaction, what seems to be her best, most preferred possible alternatives. Several models of preferences have been developed in the literature on Artificial Intelligence and Operations Research, offering the possibility to represent complex preferences over multi-attribute domains in some rather compact form. The richness of the models comes at a cost: finding the optimal alternatives is, in general, a computationally hard problem – at least NP-nard – except for some quite restrictive models.

On the other hand, the recent expansion of the use of Artificial Intelligence in many fields is accompanied by a demand for interpretability of the models used. In order to be trusted by its users, an intelligent system should not only be able to solve problems, but also capable to explain its solutions ; the explanation should allow the user to identify elements on which the solution is based.

The topic of this internship is to compare models of preferences from the point of view of their explainability. Explaining models of preferences is an emerging topic. Recent works have explored explainability of Multi-Criteria Decision Aiding models [Belahcene et al, 2016 ; Labreuche & Fossier, 2018] and of Bayesian Networks classifiers [Marques-Silva et al., NIPS 2020; Koopman & Renooij, 2021]. The aim of this internship is to study what kind of explanations one would expect for decision-aid systems based on preference models ; and to study the complexity of generating such explanations for existing preference models. Models of interest can be valued CSPs, Bayesian Network as a preference representation model, lexicographic models, CP-nets

References

Belahcene, K. , Labreuche, C. , Maudet, N. , Mousseau, V. , Ouerdane, W. Explaining robust additive utility models by sequences of preference swaps. Theory and Decision 82(2), 2017, 151–183 .

Labreuche, C. , Fossier, S. Explaining Multi-Criteria Decision Aiding Models with an Extended Shapley Value. Proc. IJCAI 2018, 331–339

Marques-Silva, J. , Gerspacher, T. , Cooper, M. , Ignatiev, A. , Narodytska, N. Explaining Naive Bayes and Other Linear Classifiers with Polynomial Time and Delay. Proc NIPS 2020.

Koopman, T. , Renooij, S. Persuasive Contrastive Explanations for Bayesian Networks. Proc. ECSQARU 2021, 229–242 .

Where, when: Toulouse (IRIT/ANITI), spring 2022 (three to six months)

Contact

Helene Fargier, *IRIT/CNRS, 3IA ANITI Senior Chair,* fargier@irit.fr
Jerome Mengin, *IRIT/UPS & 3IA ANITI,* jerome.mengin@irit.fr