Sur les explications abductives préférées pour les arbres de décision et les forêts aléatoires

Gilles Audemard*, Steve Bellart*, Louenas Bounia*, Frédéric Koriche*
Jean-Marie Lagniez*, Pierre Marquis* ** [1]

Univ. Artois, CNRS, CRIL, F-62300 Lens*
Institut universitaire de France**
nom@cril.fr,
http://www.cril.univ-artois.fr/

Résumé. Dans cet article, nous nous intéressons au calcul d'*explications abductives préférées* pour des arbres de décision et des forêts aléatoires. Nous présentons deux modèles de préférence et pour chacun d'eux, nous décrivons et évaluons un algorithme de calcul de **raisons majoritaires préférées**, où les raisons majoritaires sont des explications abductives spécifiques, adaptées aux forêts aléatoires, et qui coïncident avec les raisons suffisantes dans le cas des arbres de décision. Nous montrons expérimentalement la faisabilité de l'approche. Nous montrons aussi qu'en pratique les raisons majoritaires préférées pour une instance peuvent être beaucoup moins nombreuses que ses raisons majoritaires.

1 Introduction

Expliquer les modèles de *Machine Learning (ML)* est un enjeu important qui stimule de nombreuses recherches en IA depuis plusieurs années, dans le domaine appelé aujourd'hui « IA explicable » (XAI) (voir e.g., (Ribeiro et al., 2016, 2018; Molnar, 2020)). Dans ce papier, nous nous concentrons sur le calcul d'*explications abductives* d'instances pour les arbres de décision et les forêts aléatoires. Les explications abductives visent à préciser *pourquoi* un classeur classe une instance comme positive ou négative. Pour les modèles à base d'arbres de décision comme les forêts aléatoires ou encore les arbres améliorés (*boosted trees*), les requêtes XAI, en particulier celles qui consistent à calculer une explication abductive irredondante du classement d'une instance, sont calculatoirement difficiles (Audemard et al., 2021) : il n'existe aujourd'hui aucun algorithme polynomial pour cela et l'existence de tels algorithmes est peu vraisemblable (elle aurait pour conséquence $\mathsf{P} = \mathsf{NP}$).

Plusieurs types d'explications abductives existent selon le classeur considéré. Parmi elles, on trouve les *explications de type « impliquant premier »* (Shih et al., 2018), aussi appelées *raisons suffisantes* (Darwiche et Hirth, 2020), mais également *les raisons majoritaires* (Audemard et al., 2022b). Les raisons suffisantes sont des explications irredondantes ce qui n'est pas le cas (en général) des raisons majoritaires, qui sont des explications abductives spécifiques,

1. Ce travail a été réalisé dans le cadre de la chaire ANR d'enseignement et de recherche EXPEKCTATION (ANR-19-CHIA-0005-01).