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

Gilles Audemard, Steve Bellart, Louenas Bounia, Frederic Koriche,
Jean-Marie Lagniez, Pierre Marquis

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).

adaptées aux forêts aléatoires. Calculer une raison suffisante pour une instance donnée et une forêt aléatoire est intraitable, alors qu'il existe un algorithme en temps polynomial pour générer une raison majoritaire étant donnée une instance et une forêt aléatoire. Enfin, une instance peut posséder un nombre exponentiel de raisons suffisantes de taille minimale. Et cela vaut même si l'on ne considère que des familles de classeurs « intelligibles », comme les arbres de décision. Les dériver toutes est en général infaisable et dans tous les cas, il n'y aurait pas beaucoup de sens de présenter une multitude d'explications abductives d'une instance donnée à un utilisateur (il ne pourrait pas les appréhender toutes). Enfin, comme les explications abductives d'une instance peuvent totalement différer l'une de l'autre, fournir à l'utilisateur la première explication calculée (ou les k premières) n'est pas très satisfaisant non plus.

Le travail présenté dans ce papier repose sur deux hypothèses de recherche. Premièrement, toutes les explications d'une instance ne sont pas égales, certaines sont meilleures que d'autres. Deuxièmement, la qualité d'une explication n'est en général pas intrinsèque à l'explication, mais elle dépend de l'utilisateur à qui on la fournit. Sous ces hypothèses, notre approche consiste à exploiter *des préférences utilisateur*, pour dériver seulement des *explications préférées* des instances. Se focaliser sur des explications préférées présente deux avantages : d'une part, les explications indiquées sont meilleures (elles collent le plus possible aux préférences de l'utilisateur) et d'autre part, leur nombre peut être drastiquement plus petit que le nombre total d'explications (on peut donc parfois les donner toutes).

Dans notre étude, plusieurs modèles de préférence sont proposés. Nous partons d'un modèle simple conduisant à des préférences dichotomiques sur les explications où les explications acceptables sont seulement celles qui sont construites sur un ensemble particulier d'attributs. L'étape suivante consiste à considérer des préférences plus élaborées, qui ne sont pas dichotomiques par essence mais sont plus graduelles. Un modèle de relation de préférence cardinale est présenté. Il s'appuie sur une fonction d'utilité/coût linéaire sur les attributs, où chaque attribut a un poids et les poids sont agrégés de manière additive.

Le présent article est un résumé de l'article (en anglais) (Audemard et al., 2022a), publié dans la conférence IJCAI'22. L'article original contient des résultats additionnels, non repris ici pour des raisons d'espace.

2 Préliminaires

Arbres de décision et forêts aléatoires. Pour tout entier n, nous notons $[n] = \{1, ..., n\}$. Soit F_n l'ensemble des fonctions booléennes de $\{0,1\}^n$ à $\{0,1\}$, et soit $X_n = \{x_1, \ldots, x_n\}$ un ensemble d'*attributs* booléens. Nous nommons *instance* tout $\mathbf{x} \in \{0,1\}^n$. Chaque instance $\mathbf{x}$ est décrite par ses *caractéristiques*, i.e., les valeurs booléennes données aux attributs de X_n dans $\mathbf{x}$. Pour chaque fonction $f \in F_n$, une instance $\mathbf{x}$ est un exemple *positif* si $f(\mathbf{x}) = 1$, sinon nous disons que $\mathbf{x}$ est un exemple *négatif*.

f est vue comme une formule propositionnelle quand nous l'écrivons avec les connecteurs booléens $\wedge$ (conjonction), $\vee$ (disjonction) ou $\neg$ (négation) et en utilisant les constantes 1 (vrai) et 0 (faux). f est *satisfaisable* quand f n'est pas équivalente à 0. Un *littéral* ℓ_i sur X_n décrit une variable booléenne x_i ou sa négation $\neg x_i$, également notée $\overline{x_i}$. Un *terme* t sur X_n est une conjonction de littéraux sur X_n et une *clause* c sur X_n est une disjonction de littéraux sur X_n. Souvent, t et c sont aussi vus comme des ensembles de littéraux. En particulier, si t est un terme et $S \subseteq X_n$, $t[S]$ désigne le terme dont les littéraux sont ceux de t qui sont

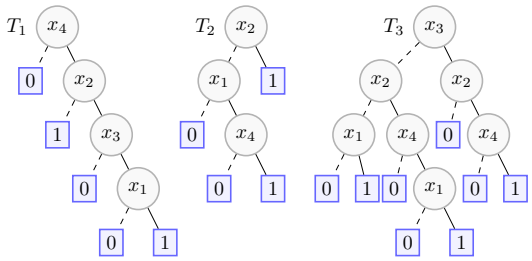

FIG. 1 – *Une forêt aléatoire pour reconnaître les orchidées Cattleya. Le fils gauche (resp. droit) d'un nœud de décision étiquetté par x_i correspond à l'affectation de x_i à 0 (resp. 1).*

sur S. Une formule DNF (forme normale disjonctive) est une disjonction de *termes* et une formule CNF (forme normale conjonctive) est une conjonction de *clauses*. Un terme t *couvre* une instance $\mathbf{z}$ si l'ensemble des littéraux composant t est inclus dans l'ensemble des littéraux composant le terme représentant z, noté $t_{\mathbf{z}}$. Un *impliquant* d'une formule f est un terme t tel que toute instance couverte par t est un exemple positif de f. Un *impliquant premier* t d'une formule f est un impliquant de f tel que tout terme formé comme la conjonction d'un sous-ensemble propre des littéraux de t n'est pas un impliquant de f.

Parmi les représentations de fonctions booléennes, on trouve les arbres de décision et les forêts aléatoires. Un *arbre de décision* (booléen) sur X_n est un arbre binaire T dont chaque nœud interne correspond à une des variables d'entrée et dont chaque feuille est étiquetée soit par 0, soit par 1. Chaque variable est supposée n'apparaître qu'une fois dans chaque chemin racine/feuille de l'arbre. La valeur de $T(\mathbf{x}) \in \{0,1\}$ de T pour une instance $\mathbf{x}$ est donnée par la valeur de la feuille atteinte en partant de la racine : à chaque nœud, si la variable $\mathbf{x}_i$ associée au nœud considéré vaut 1 alors nous continuons sur le fils droit, sinon sur le fils gauche. Une *forêt aléatoire* (booléenne) sur X_n est un ensemble $F = \{T_1, \ldots, T_m\}$, où chaque $T_i (i \in [m])$ est un arbre de décision sur X_n et tel que la valeur de $F(\mathbf{x}) \in \{0,1\}$ d'une instance $\mathbf{x}$ est donnée par :

$$F(\mathbf{x}) = \begin{cases} 1 & \text{si } \frac{1}{m}\sum_{i=1}^{m} T_i(\mathbf{x}) > \frac{1}{2} \\ 0 & \text{sinon} \end{cases}$$

La taille de F est donnée par $|F| = \Sigma_{i=1}^{m}|T_i|$, où $|T_i|$ est le nombre de nœuds apparaissant dans T_i. L'ensemble des arbres de décision définis sur X_n est noté DT_n et l'ensemble des forêts aléatoires définies sur X_n et possédant au moins m arbres est noté $RF_{n,m}$. De plus, on note RF_n l'union des $RF_{n,m}$ pour tout $m \in \mathbb{N}$.

Exemple 1. *La forêt aléatoire $F = \{T_1, T_2, T_3\}$ donnée à la figure 1 et reprise de (Audemard et al., 2022b) est formée de trois arbres de décision. Cette forêt sépare les orchidées Cattleya des autres orchidées en s'appuyant sur quatre attributs : x_1 : "a des fleurs parfumées", x_2 : "a une ou deux feuilles", x_3 : "a des fleurs de grande taille" et x_4 : "est sympodiale".*

Explications abductives. Formellement, pour $f \in F_n$ et $\mathbf{x} \in \{0,1\}^n$, une *explication abductive* (Ignatiev et al., 2019) (ou *raison*) pour $\mathbf{x}$ étant donnée f est un impliquant t de f (ou de $\neg f$ dans le cas où $f(\mathbf{x}) = 0$) qui couvre $\mathbf{x}$. Il existe toujours une explication abductive t de

$\mathbf{x}$ étant donnée f car $t = t_{\mathbf{x}}$ est trivialement une telle explication. Ce faisant, nous allons dans le reste de cette section, nous concentrer sur des formes plus concises d'explication abductive.

Définition 1. *Soient $F \in RF_n$ une forêt aléatoire et $\mathbf{x} \in \{0,1\}^n$ une instance telle que $F(\mathbf{x}) = 1$ (resp. $F(\mathbf{x}) = 0$). Une* raison suffisante *pour $\mathbf{x}$ étant donnée F est un impliquant premier t de F (resp. de $\neg F$) tel que t couvre $\mathbf{x}$.*

Déterminer si un terme donné est une raison suffisante pour une instance $\mathbf{x}$ étant donnée une forêt aléatoire F a été montré DP-complet (Izza et Marques-Silva, 2021).

Introduites dans (Audemard et al., 2022b), les *raisons majoritaires* sont des explications abductives spécifiques aux forêts aléatoires.

Définition 2. *Soient $F = \{T_1, \ldots, T_m\}$ une forêt aléatoire de $RF_{n,m}$ et $\mathbf{x} \in \{0,1\}^n$ une instance telle que $F(\mathbf{x}) = 1$ (resp. $F(\mathbf{x}) = 0$). Une* raison majoritaire *pour $\mathbf{x}$ étant donnée F est un terme t couvrant $\mathbf{x}$ qui est un impliquant d'au moins $\lfloor \frac{m}{2} \rfloor + 1$ arbres de décision T_i (resp. $\neg T_i$) et $\forall l \in t, t \backslash \{l\}$ ne satisfait plus la condition précédente.*

En général, les raisons majoritaires peuvent contenir des caractéristiques redondantes. Toutefois, pour les forêts ne comportant qu'un seul arbre, les raisons majoritaires pour une instance coïncident avec ses raisons suffisantes.

Reprenons l'exemple 1 et considérons l'instance $\boldsymbol{x} = (1,1,1,1)$. Puisque $F(\boldsymbol{x}) = 1$, $\boldsymbol{x}$ est reconnue comme une orchidée *Cattleya*. Ce classement de $\boldsymbol{x}$ par F peut être expliqué de plusieurs manières. Ainsi, $x_2 \wedge x_3 \wedge x_4$ et $x_1 \wedge x_4$ sont les raisons suffisantes pour $\boldsymbol{x}$ étant donnée F. Les raisons majoritaires pour $\boldsymbol{x}$ étant donnée F sont $x_1 \wedge x_2 \wedge x_4$, $x_1 \wedge x_3 \wedge x_4$ et $x_2 \wedge x_3 \wedge x_4$. Les raisons majoritaires $x_1 \wedge x_2 \wedge x_4$ et $x_1 \wedge x_3 \wedge x_4$ contiennent des caractéristiques redondantes (x_2 pour la première et x_3 pour la seconde).

3 Explications abductives préférées

Définir *in abstracto* ce qu'est une explication « préférée » ou, au minimum, « suffisamment bonne » est difficile en général. Il n'y a pas de consensus à ce sujet et une dépendance claire à l'utilisateur, voir par exemple (Doshi-Velez et Kim, 2017; Lipton, 2018).

3.1 Préférences dichotomiques sur les explications

Nous présentons d'abord un modèle où les préférences de l'utilisateur sont *dichotomiques*, c'est-à-dire que les explications peuvent être partitionées en deux ensembles : l'une contenant des raisons **« suffisamment bonnes »** et l'autre contenant des raisons jugées **« pas assez bonnes »**.

Définition 3. *Soient $f \in \mathcal{F}_n$, $S \subseteq X_n$ et $\boldsymbol{x} \in \{0,1\}^n$. Une* explication abductive basée sur *S pour $\boldsymbol{x}$ étant donnée f est une explication abductive t pour $\boldsymbol{x}$ étant donnée f telle que $Var(t) \subseteq S$.*

Ce modèle s'appuie sur un sous-ensemble $S \subseteq X_n$ et permet de gérer quelques situations d'intérêt où l'utilisateur veut écarter les explications qui font référence à des *concepts non compréhensibles* (modélisés comme des attributs en dehors de S). De telles explications à rejeter

peuvent faire référence à des attributs correspondant à des notions trop techniques, qui ne sont pas comprises par l'utilisateur (par exemple, un terme médical pour un patient qui n'est pas médecin), ou encore parce qu'il est ne sont pas documentés ou sont assez vagues par essence. Ainsi, sur l'exemple 1, l'utilisateur voudrait pouvoir calculer une explication abductive pour $\boldsymbol{x} = (1, 1, 1, 1)$ qui ne fait pas intervenir l'attribut x_4 ("est sympodiale") car il ne connaît pas le sens de cet adjectif (mais c'est impossible ici : x_4 est un attribut nécessaire à l'explication).

S'assurer que seules les explications basées sur les attributs de S sont générées est également utile pour atteindre d'autres objectifs. Ainsi, la présence de certains *éléments protégés* doit être évitée dans les explications chaque fois que cela est possible, car l'impossibilité de laisser ces attributs de côté reflète précisément le fait que la décision prise était biaisée (Darwiche et Hirth, 2020). Par exemple, considérons dans une procédure d'admission à l'université, un candidat pour lequel la décision prise par le classeur est positive : si toute explication abductive de cette décision mentionne le fait que le candidat vient d'une **ville natale riche (élément protégé)**, la décision est biaisée. Par conséquent, « venant d'une ville natale riche » ne devrait pas appartenir à S. Au-delà des problèmes de compréhensibilité ou de biais, la présence d'attributs *non actionnables* doit être évitée dans les explications. Ne pas être actionnable signifie simplement que l'utilisateur ne peut pas (ou peut difficilement) changer la valeur de cet attribut dans l'instance pour laquelle une explication est recherchée.

Pour calculer une raison majoritaire pour une instance $\mathbf{x}$ étant donnée une forêt aléatoire F, on peut utiliser *un algorithme glouton* (voir (Audemard et al., 2022b) pour plus de détails) : partant d'un terme t qui couvre l'instance $\boldsymbol{x}$ à traiter (typiquement $t = t_{\boldsymbol{x}}$), on parcourt les caractéristiques ℓ de t et on supprime ℓ de t chaque fois que le terme obtenu reste un impliquant d'au moins $\lfloor \frac{m}{2} \rfloor + 1$ arbres de décision T_i (resp. $\neg T_i$) lorsque $F(\mathbf{x}) = 1$ (resp. $F(\mathbf{x}) = 0$). Il est facile d'étendre un tel algorithme pour décider s'il existe une raison majoritaire basée sur un ensemble prédéfini S d'attributs.

Proposition 1. *Soient $F \in RF_n$ et $\boldsymbol{x} \in \{0,1\}^n$. Pour un ensemble $S \subseteq X_n$, décider si une raison majoritaire basée sur S pour $\boldsymbol{x}$ étant donnée F existe, et dériver une telle raison lorsque c'est le cas, peut être réalisé en temps $\mathcal{O}(n|F|)$ en utilisant un algorithme glouton.*

3.2 Préférences graduelles sur les explications

Le modèle précédent induit des préférences dichotomiques sur les explications. Si une telle séparation en deux classes est commode dans certains cas, on s'attend à *plus de gradualité* dans d'autres cas, afin (par exemple) d'éviter la présence de certains attributs dans les explications sans l'interdire pour autant.

Pour cela, une approche consiste à tirer parti d'une *fonction d'utilité* (ou d'une fonction de coût). Dans notre cadre, on associe une valeur de disutilité (un poids représentant un coût) à chaque attribut : plus le poids est élevé, moins l'attribut est intéressant. Cette fois, la relation de préférence résultante est un pré-ordre total sur les explications, les meilleures explications étant celles de coût minimal.

Définition 4. *Soit $f \in \mathcal{F}_n$. Soit $w : X_n \to \mathbb{N}^*$ un vecteur de poids, où un poids est associé à chaque attribut. Une* explication abductive de poids minimal *pour $\boldsymbol{x}$ étant donné f et w est une explication abductive t pour $\boldsymbol{x}$ et f minimisant $\Sigma_{x \in Var(t)} w(x)$.*

Afin de calculer une raison majoritaire de poids minimal, contrairement à ce qui était possible pour les autres modèles de préférence, on ne peut tirer parti d'aucune variante en temps

polynomial de l'algorithme glouton pour calculer une raison majoritaire. En effet, dans le cas général, le calcul d'une raison majoritaire de poids minimal est NP-difficile. Cela vient du fait que dériver une raison majoritaire de taille minimale pour une instance $\boldsymbol{x}$ étant donnée une forêt aléatoire F est NP-difficile (Audemard et al., 2022b) puisqu'une raison majoritaire de taille minimale pour $\boldsymbol{x}$ étant donnée F est une raison majoritaire de poids minimal pour pour $\boldsymbol{x}$ étant donnée F et l'application w_1 uniformément égale à 1 (i.e., $\forall i \in [n]$, $w_1(x_i) = 1$).

Néanmoins, on peut généraliser l'approche présentée dans (Audemard et al., 2022b) pour calculer des raisons majoritaires de taille minimale au cas des raisons majoritaires de poids minimal. Formellement, on réduit le problème du calcul d'une explication de poids minimal au problème WEIGHTED PARTIAL MAXSAT (voir par exemple Li et Manyà (2009)). Une instance de WEIGHTED PARTIAL MAXSAT est la donnée d'un ensemble C_{hard} de clauses « dures« (à satisfaire) et d'un ensemble C_{soft} de clauses « souples » pondérées par des entiers positifs. Une solution d'une telle instance est une interprétation qui satisfait toutes les clauses de C_{hard} et maximise la somme des poids des clauses de C_{soft} qui sont satisfaites.

Proposition 2. *Soient $F = \{T_1, \ldots, T_m\}$ une forêt aléatoire dans $RF_{n,m}$ et une instance $\boldsymbol{x} \in \{0,1\}^n$ telle que $F(\boldsymbol{x}) = 1$.*[2] *Soit $w : X_n \to \mathbb{N}^*$ une application. Une raison majoritaire de poids minimal pour $\boldsymbol{x}$ étant donné F et w est donnée par $t_{\boldsymbol{x}} \cap t_{\boldsymbol{v}^*}$, où $\boldsymbol{v}^*$ est une solution de l'instance $(C_{\text{soft}}, C_{\text{hard}})$ du problème* WEIGHTED PARTIAL MAXSAT *telle que $C_{\text{soft}} = \{(\overline{x_i}, w(x_i)) : x_i \in t_{\boldsymbol{x}}\} \cup \{(x_i, w(x_i)) : \overline{x}_i \in t_{\boldsymbol{x}}\}$, $C_{\text{hard}} = \{(\overline{y_j} \vee c[\boldsymbol{x}], \infty) : i \in [m], c \in \texttt{CNF}(T_i)\} \cup \texttt{CNF}(\sum_{i=1}^m y_i > \frac{m}{2})$ où $y_1, \ldots, y_m$ sont des variables auxiliaires, et $\texttt{CNF}(\sum_{i=1}^m y_i > \frac{m}{2})$ est un encodage* CNF *de la contrainte de cardinalité $\sum_{i=1}^m y_i > \frac{m}{2}$.*

Sur l'exemple 1, si l'utilisateur souhaite obtenir une raison majoritaire pour $\boldsymbol{x} = (1,1,1,1)$ étant donnée F qui fait intervenir autant que possible les propriétés des fleurs (x_1 et x_3) plus que des feuilles (x_2) et pas celle portant sur la tige (x_4), il peut considérer (par exemple) w telle que $w(x_1) = w(x_3) = 1$, $w(x_2) = 3$ et $w(x_4) = 6$. La raison majoritaire (unique) de poids minimal pour $\boldsymbol{x} = (1,1,1,1)$ étant donnée F et w, qui vaut $x_1 \wedge x_3 \wedge x_4$ et qui a pour poids 8, pourra alors être calculée.

4 Expérimentations

Protocole expérimental. Nous avons sélectionné 22 datasets disponibles sur Kaggle (`www.kaggle.com`), OpenML (`www.openml.org`) ou UCI (`archive.ics.uci.edu/ml`). Pour chaque dataset, nous avons appris des forêts aléatoires et effectué une validation croisée à 10 blocs et pour un sous-ensemble d'au plus 250 instances. Ces datasets n'étant pas nativement associés à des préférences d'utilisateurs, nous avons opté pour des fonctions de poids reflétant des préférences possibles de différents types : la fonction w_1 uniformément égale à 1, les valeurs de **SHAP** et **f-importance** des attributs (qui se veulent des mesures globales de l'importance des différents attributs dans les classements effectués) et enfin **wordfreq** qui est la fréquence du substantif dénotant l'attribut en langue anglaise (et peut être vu comme un indicateur de son intelligibilité).

Nous avons calculé et énuméré les raisons majoritaires de poids minimal des instances en utilisant ces différentes fonctions de poids. Pour l'autre modèle de préférence présenté dans le

2. Si $F(\boldsymbol{x}) = 1$, on considérera la forêt aléatoire $\neg F$ au lieu de F ; elle se calcule en temps linéaire à partir de F, voir (Audemard et al., 2022b).

dataset / random forest				minimum-size			SHAP			f-importance			wordfreq		
name	%A	#B	#I	1	A	nb	1	A	nb	1	A	nb	1	A	nb
divorce	97.65	50	170	170	161	41.6 (±77.4)	169	169	1.2 (±0.4)	170	170	1.1 (±0.3)	170	170	1.0 (±0.1)
compas	66.51	65	250	250	250	6.0 (±9.9)	249	249	1.9 (±1.8)	247	243	2.7 (±3.9)	250	250	2.4 (±2.7)
employee	83.17	72	250	243	174	8.9 (±12.6)	243	235	2.0 (±2.1)	249	245	2.1 (±3.5)	239	204	4.4 (±7.6)
student mat	90.63	144	250	250	217	44.7 (±60.4)	250	250	1.1 (±0.2)	250	250	1.1 (±0.3)	250	250	1.2 (±0.4)
student por	91.99	171	250	19	10	43.0 (±38.4)	16	16	1.1 (±0.3)	14	14	1.4 (±0.8)	25	24	1.4 (±0.7)
anneal 2	99.11	203	250	250	200	26.8 (±39.0)	240	240	1.1 (±0.3)	241	240	1.1 (±0.3)	248	248	1.1 (±0.4)
placement	93.55	262	215	215	145	50.7 (±61.0)	215	215	1.4 (±1.3)	215	213	1.3 (±0.7)	215	212	1.2 (±0.6)
heart	78.3	263	250	250	236	41.7 (±59.5)	250	250	1.3 (±0.6)	250	250	1.3 (±0.6)	250	250	1.4 (±0.9)
diabetes	72.28	433	250	250	248	18.2 (±37.7)	250	250	1.3 (±1.1)	250	250	1.3 (±0.7)	250	250	1.1 (±0.4)
horse	87.31	540	250	62	6	72.7 (±46.0)	50	50	1.4 (±0.8)	55	55	1.4 (±0.8)	42	41	1.4 (±0.6)
ind. l. pat.	69.61	613	250	250	187	54.1 (±54.3)	250	250	1.6 (±1.7)	250	250	1.5 (±0.9)	250	250	2.3 (±2.8)
banknote	99.42	652	250	190	37	11.1 (±7.1)	215	207	2.0 (±2.7)	237	231	1.8 (±1.8)	160	150	1.2 (±0.5)
startup	80.18	3517	250	57	0	- (-)	40	38	1.4 (±0.7)	43	42	1.5 (±0.9)	43	38	1.8 (±1.1)
farm-ads	87.3	5389	250	25	0	- (-)	250	250	1.0 (±0.0)	11	11	1.0 (±0.0)	25	25	1.2 (±0.4)

papier (et les deux autres modèles décrits dans (Audemard et al., 2022a)), nous n'avons pas réalisé d'expérimentations car la question de la faisabilité pratique ne se posait pas : les algorithmes de génération d'une explication proposés sont tous en temps polynomial. La plupart du temps, les poids des attributs n'étant pas des entiers positifs, une transformation affine de la fonction de poids a été réalisée (l'ordre induit sur les explications est préservé de sorte que les raisons majoritaires de poids minimal ne changent pas).

Nous avons considéré un temps limite de génération de 60 secondes par instance. Étant donné que les fonctions de poids font partie de l'entrée, nous n'avons pas compté le temps de calcul nécessaire pour les générer dans les 60 secondes. Nous avons utilisé le solveur WEIGHTED PARTIAL MAXSAT `openwbo` (Martins et al., 2014) pour calculer les raisons majoritaires de poids minimal. Tous les calculs ont été réalisés sur un ordinateur équipé de processeur Intel(R) Core(TM) i9-9900 à 3,10 GHz - 16 cœurs et 64 Go de mémoire

Résultats expérimentaux. Le tableau ci-avant présente un extrait des résultats obtenus, basé sur 14 datasets. (%A) est la précision moyenne des forêts, (#B) le nombre moyen d'attributs booléens dans celles-ci et (#I) le nombre d'instances. Pour chaque fonction de poids, (1) (resp. (A)) donne le nombre d'instances du dataset pour lesquelles au moins une (resp. toutes les) raison(s) majoritaire(s) préférée(s) ont été calculées en 60 secondes ; la colonne (nb) donne la moyenne (et l'écart-type) du nombre de raisons majoritaires obtenues lorsqu'elles ont toutes été calculées. Les résultats empiriques obtenus montrent clairement que le calcul de raisons majoritaires préférées est très souvent faisable en pratique.

5 Conclusion

Dans cet article, nous avons considéré le problème de la génération d'explications abductives du classement d'instances, et en particulier du calcul de raisons majoritaires, quand les classeurs utilisés sont des arbres de décision ou des forêts aléatoires. Comme les raisons majoritaires peuvent être en nombre exponentiel, il est en général hors de portée de les calculer toutes. Une approche pour pallier ce problème consiste à définir des modèles de préférence et à les exploiter pour dériver seulement des explications préférées. Dans les sections précédentes, nous avons présenté deux modèles de ce type et décrit des algorithmes de génération d'explications préférées. Nos expérimentations ont montré la faisabilité de cette génération. Elles ont aussi mis en évidence que l'exploitation des préférences de l'utilisateur peut réduire considérablement le nombre de raisons, rendant leur énumération possible dans des situations où le calcul de toutes les raisons majoritaires ne le serait pas.

Références

Audemard, G., S. Bellart, L. Bounia, F. Koriche, J. Lagniez, et P. Marquis (2021). On the computational intelligibility of boolean classifiers. In *Proc. of KR'21*, pp. 74–86.

Audemard, G., S. Bellart, L. Bounia, F. Koriche, J. Lagniez, et P. Marquis (2022a). On preferred abductive explanations for decision trees and random forests. In *Proc. of IJCAI'22*, pp. 634–650.

Audemard, G., S. Bellart, L. Bounia, F. Koriche, J.-M. Lagniez, et P. Marquis (2022b). Trading complexity for sparsity in random forest explanations. In *Proc. of AAAI'22*, pp. 5461–5469.

Darwiche, A. et A. Hirth (2020). On the reasons behind decisions. In *Proc. of ECAI'20*, pp. 712–720.

Doshi-Velez, F. et B. Kim (2017). A roadmap for a rigorous science of interpretability. *CoRR abs/1702.08608*.

Ignatiev, A., N. Narodytska, et J. Marques-Silva (2019). Abduction-based explanations for machine learning models. In *Proc. of AAAI'19*, pp. 1511–1519.

Izza, Y. et J. Marques-Silva (2021). On explaining random forests with SAT. In *Proc. of IJCAI'21*, pp. 2584–2591.

Li, C. M. et F. Manyà (2009). Maxsat, hard and soft constraints. In *Handbook of Satisfiability*, pp. 613–631. IOS Press.

Lipton, Z. C. (2018). The mythos of model interpretability. *CACM 61*(10), 36–43.

Martins, R., V. M. Manquinho, et I. Lynce (2014). Open-WBO : A modular MaxSAT solver. In *Proc. of SAT'14*, pp. 438–445.

Molnar, C. (2020). *Interpretable Machine Learning*. Leanpub.

Ribeiro, M. T., S. Singh, et C. Guestrin (2016). "why should I trust you ?" : Explaining the predictions of any classifier. In *Proc. of SIGKDD'16*, pp. 1135–1144.

Ribeiro, M. T., S. Singh, et C. Guestrin (2018). Anchors : High-precision model-agnostic explanations. In *Proc. of AAAI'18*, pp. 1527–1535.

Shih, A., A. Choi, et A. Darwiche (2018). A symbolic approach to explaining bayesian network classifiers. In *Proc. of IJCAI'18*, pp. 5103–5111.

Summary

In this paper, we are interested in computing *preferred abductive explanations* for decision trees and random forests. We present two preference models and for each of them, we describe and evaluate an algorithm for computing **preferred majoritary reasons**, where majoritary reasons are specific abductive explanations, suited to random forests, and which coincide with sufficient reasons in the case of decision trees. We experimentally show the feasibility of the approach. We also show that in practice the preferred majoritary reasons for an instance can be much less numerous than its majoritary reasons.