Sur le pouvoir explicatif des arbres de décision

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

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

Résumé. Les arbres de décision constituent un modèle d'apprentissage adapté aux applications pour lesquelles l'interprétabilité des décisions est d'une importance primordiale. Nous examinons ici la capacité des arbres de décision binaires à extraire, minimiser et compter des explications abductives / contrastives. Nous montrons que l'ensemble de toutes les explications abductives irredondantes (ou raisons suffisantes) d'une instance peut être de taille exponentielle. Aussi, générer l'intégralité de cet ensemble peut se révéler hors de portée. De plus, deux raisons suffisantes d'une même instance peuvent différer sur tous leurs attributs. Ainsi, le calcul d'une seule raison suffisante ne donne qu'une vision parcellaire des explications abductives possibles. Nous introduisons les notions d'attribut nécessaire / pertinent pour l'explication et la notion d'importance explicative d'un attribut et nous montrons que ces notions peuvent être utiles pour dériver une vue synthétique des raisons suffisantes d'une instance.

1 Introduction

Expliquer une décision à une personne, c'est donner les détails ou les raisons qui aident cette personne à comprendre pourquoi une telle décision a été prise. Lorsque les décisions sont prises par des modèles d'apprentissage automatique (ML) dits opaques, comme les forêts aléatoires, les SVM et les réseaux de neurones, la génération d'explications est une tâche complexe. Pour autant, avec le nombre croissant d'applications qui reposent sur des techniques d'apprentissage automatique, les recherches sur l'IA explicable (XAI) sont devenues essentielles. Elles visent à développer des méthodes et des approches efficaces pour interpréter les modèles d'apprentissage et expliquer les décisions prises (Frosst et Hinton, 2017; Guidotti et al., 2019; Hooker et al., 2019; Huysmans et al., 2011; Ignatiev et al., 2019; Lundberg et Lee, 2017; Miller, 2019; Molnar, 2019; Ribeiro et al., 2016; Shih et al., 2019).

Dans cet article, nous nous intéressons aux classeurs booléens où seules deux décisions (classements) sont possibles, 1 pour les instances positives, et 0 pour les instances négatives. Peu importe que l'instance considérée $\boldsymbol{x}$ soit positive ou pas, la recherche d'explications de son classement est une question importante dans bien des cas (Miller, 2019). D'une part, les explications « abductives » visent à expliquer pourquoi $\boldsymbol{x}$ est classée comme elle a été classée

par le modèle d'apprentissage automatique (répondant ainsi à la question « pourquoi ce classement ? »). D'autre part, les explications « contrastives », également appelées « contrefactuelles », ont pour but d'expliquer pourquoi $\boldsymbol{x}$ n'a pas été classée par le modèle comme on l'attendait (répondant ainsi à la question « pourquoi pas l'autre classement ? »). Dans les deux cas, des explications aussi simples que possible sont privilégiées.

Bien qu'il n'y ait pas de définition exacte ou de notion formelle de ce que couvre l'interprétabilité (Lipton, 2018), les arbres de décision (Breiman et al., 1984; Quinlan, 1986) sont très largement considérés comme parmi les modèles d'apprentissage automatique les plus interprétables pour les problèmes de classement. Pour cette raison, ils sont souvent considérés comme des modèles cibles pour distiller un modèle opaque en un modèle compréhensible (Breiman et Shang, 1996; Frosst et Hinton, 2017). De plus, les arbres de décision sont aussi utilisés comme éléments primitifs à la construction d'autres modèles qui sont moins interprétables, mais potentiellement plus précis, à l'image des forêts aléatoires (RF) (Breiman, 2001). L'interprétabilité des arbres de décision repose sur deux caractéristiques clés. D'une part, les arbres de décision sont *transparents* : chaque nœud dans un arbre de décision a une certaine signification, et les principes utilisés pour générer les nœuds peuvent être expliqués simplement. D'autre part, les arbres de décision sont *localement explicables*. En effet, on peut facilement extraire une raison appelée « directe » à partir de l'instance à classer. Une telle raison est une explication abductive de cette instance. Toutefois, les raisons directes peuvent contenir de nombreux attributs redondants (Izza et al., 2020). Cela motive à prendre en compte d'autres types d'explications abductives pour les arbres de décision, à savoir, les raisons suffisantes (Darwiche et Hirth, 2020) (appelées également PI-explications (Shih et al., 2018)), qui sont des explications abductives non redondantes, et les raisons suffisantes minimales (c.-à-d. des raisons suffisantes de taille minimale).

Dans cet article, nous examinons la capacité des arbres de décision binaires à dériver, minimiser et compter des raisons suffisantes et des explications contrastives. Nous montrons que l'ensemble de toutes les raisons suffisantes de taille minimale pour une instance $\boldsymbol{x}$ étant donné un arbre de décision T peut être exponentiellement plus grand que la taille de l'entrée. Quand c'est le cas, générer l'ensemble de toutes les raisons suffisantes (c-à-d la raison complète de l'instance (Darwiche et Hirth, 2020)) peut se relever hors de portée. De plus, extraire une seule raison suffisante ne donne qu'une vision parcellaire de l'explication. En effet, deux raisons suffisantes pour une même instance peuvent différer sur tous leurs attributs. Pour traiter cette question et générer des explications concises de l'ensemble de toutes les raisons suffisantes, nous introduisons les notions d'attribut pertinent et d'attribut nécessaire qui caractérisent respectivement les attributs qui apparaissent dans au moins une et dans toutes les raisons suffisantes, et nous montrons que nous pouvons les calculer en temps polynomial. Nous introduisons également la notion de pouvoir explicatif d'un attribut, qui donne la fréquence de chaque attribut dans l'ensemble de toutes les raisons suffisantes. Nous montrons comment calculer le pouvoir explicatif d'un attribut à l'aide d'une opération de comptage de modèles. Nous expliquons également comment énumérer des raisons suffisantes de taille minimale, ce qui permet de les compter quand elles ne sont pas trop nombreuses. Enfin, nous montrons que le calcul de l'ensemble des explications contrastives d'une instance peut être réalisé en temps polynomial. Les preuves des propositions ainsi que le code utilisé dans les expérimentations et des résultats empiriques plus détaillés sont disponibles à partir de `www.cril.univ-artois.fr/expekctation/papers.html`.

2 Arbres de décision, explications abductive et contrastive

Pour un entier n, soit $[n]$ l'ensemble $\{1, \cdots, n\}$. nous notons $\mathcal{F}_n$ l'ensemble des fonctions booléennes de $\{0,1\}^n$ dans $\{0,1\}$ et $X_n = \{x_1, \cdots, x_n\}$ l'ensemble de variables booléennes d'entrée. Toute assignation (affectation) $\boldsymbol{x} \in X_n$ est appelée une *instance*. Si $f(\boldsymbol{x}) = 1$ pour $f \in \mathcal{F}_n$, alors $\boldsymbol{x}$ est appelée *modèle* de f. L'instance $\boldsymbol{x}$ est *positive* si $f(\boldsymbol{x}) = 1$ et *négative* si $f(\boldsymbol{x}) = 0$.

On se réfère à f comme une *formule propositionnelle* quand f est représentée à l'aide des connecteurs, $\wedge$ (conjonction), $\vee$ (disjonction), et $\neg$ (négation), et en utilisant éventuellement les constantes booléennes 1 (vrai) et 0 (faux). Un *littéral* ℓ est une variable x_i (un littéral positif) ou sa négation $\neg x_i$, également notée $\overline{x}_i$ (un littéral négatif). La polarité d'une variable x_i dans le littéral x_i est positive et elle est négative dans le littéral $\neg x_i$.

Un *terme* (ou *monôme*) t est une conjonction de littéraux, et une *clause* c est une disjonction de littéraux. Une *formule* DNF est une disjonction de termes et une *formule* CNF est une conjonction de clauses. $Var(f)$ est l'ensemble des variables d'une formule f. Une formule f est cohérente si et seulement si elle a un modèle. Une formule CNF est monotone si chaque variable qui la compose apparaît toujours avec la même polarité. Une formule f_1 *implique* une formule f_2, noté $f_1 \models f_2$, si et seulement si chaque modèle de f_1 est un modèle de f_2. Deux formules f_1 et f_2 sont *équivalentes*, noté $f_1 \equiv f_2$ lorsqu'elles ont les mêmes modèles.

Le *conditionnement* d'une formule f par un littéral ℓ, noté $f \mid \ell$, est la formule obtenue de f en substituant à chaque occurrence de x_i la constante 1 (resp. 0) et à chaque occurrence de $\overline{x}_i$ la constante 0 (resp. 1) si $\ell = x_i$ (resp. $\ell = \overline{x}_i$). Dans ce qui suit, nous considérons souvent les affectations comme des termes, et les termes et clauses comme des ensembles de littéraux. Étant donnée une affectation $z \in \{0,1\}^n$, le terme correspondant $t_{\boldsymbol{z}}$ est défini comme suit : $t_{\boldsymbol{z}} = \bigwedge_{i=1}^{n} x_i^{z_i}$ où $x_i^0 = \overline{x}_i$ et $x_i^1 = x_i$. Un terme t *couvre* une affectation z si $t \subseteq t_z$. Un *impliquant* d'une fonction booléenne f est un terme qui implique f. Un *impliquant premier* de f est un impliquant t de f tel qu'aucun sous-ensemble de t n'est un impliquant de f. Il s'agit donc d'un impliquant irredondant de f (tous ses littéraux sont pertinents). Rappelons maintenant le modèle des arbres de décision.

Définition 1 (Arbre de décision)**.** *Un arbre de décision booléen sur X_n est un arbre de décision binaire, où chaque nœud interne est étiqueté par l'une des n variables booléennes d'entrée, et pour lequel chacune des feuilles est étiquetée par 0 ou 1. Chaque variable apparaît au plus une fois sur n'importe quel chemin de la racine à feuille. La valeur $T(\boldsymbol{x}) \in \{0,1\}$ de T pour l'instance d'entrée $\boldsymbol{x}$ est donnée par l'étiquette de la feuille atteinte depuis la racine comme suit : à chaque nœud, on suit le fils gauche ou droit selon que la valeur d'entrée de la variable correspondante est 0 ou 1. La taille de T, noté $|T|$ est donnée par le nombre de ses nœuds.*

La classe des arbres de décision sur X_n est notée $\mathtt{DT}_n$. Il est bien connu que tout arbre $T \in \mathtt{DT}_n$ peut se transformer en une disjonction de termes équivalent en temps linéaire, notée $\mathtt{DNF}(T)$, chaque terme correspondant à un chemin de la racine à une feuille 1. De même, T peut se transformer en temps linéaire en une conjonction de clauses, notée $\mathtt{CNF}(T)$, où chaque clause est la négation d'un terme correspondant à un chemin de la racine à une feuille 0.

L'exemple qui suit, utilisé tout au long de l'article, illustre le concept d'arbre de décision.

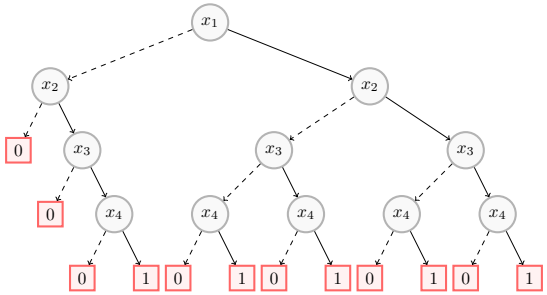

FIG. 1: Un arbre de décision T pour reconnaître les orchidées *Cattleya*.

Exemple 1. *L'arbre de décision de la figure 1 sépare les orchidées Cattleya des autres orchidées en utilisant les attributs suivants :* x_1 *: « a des fleurs parfumées »,* x_2 *: « a une ou deux feuilles »,* x_3 *: « a de larges fleurs »et* x_4 *: « est sympodiale ».*

Définition 2 (Raison directe)**.** *Soient un arbre de décision* $T \in \mathtt{DT}_n$ *et une instance* $\boldsymbol{x} \in \{0,1\}^n$*. La* raison directe *de* $\boldsymbol{x}$ *étant donné* T *est le terme, noté* $t_{\boldsymbol{x}}^T$*, correspondant au chemin unique de la racine à la feuille de* T *qui est compatible avec* $\boldsymbol{x}$*.*

Les raisons directes sont des explications abductives des instances. D'autres explications abductives sont données par les *raisons suffisantes* (Darwiche et Hirth, 2020). Elles différent en général des raisons directes et ne sont pas spécifiques aux arbres de décision :

Définition 3 (Raison suffisante)**.** *Soient* $f \in \mathcal{F}_n$ *et* $\boldsymbol{x} \in \{0,1\}^n$ *tel que* $f(\boldsymbol{x}) = 1$ *(resp.* $f(\boldsymbol{x}) = 0$*). Une* raison suffisante *de* $\boldsymbol{x}$ *étant donnée* f *est un impliquant premier* t *de* f *(resp.* $\neg f$*) qui couvre* $\boldsymbol{x}$*.* $sr(\boldsymbol{x}, f)$ *désigne l'ensemble des raisons suffisantes de* $\boldsymbol{x}$ *étant donnée* f*.*

Une raison suffisante (Darwiche et Hirth, 2020) (ou PI-explication (Shih et al., 2018)) d'une instance $\boldsymbol{x}$ étant donnée une fonction booléenne f est un sous-ensemble t de $\boldsymbol{x}$ qui est minimal par rapport à l'inclusion d'ensemble et tel que toute instance $\boldsymbol{x}'$ qui partage l'ensemble t est classée par f comme $\boldsymbol{x}$. Ainsi, quand t couvre x, lorsque $f(\boldsymbol{x}) = 1$, t est une raison suffisante de $\boldsymbol{x}$ étant donnée f si et seulement si t est un impliquant premier de f, et lorsque $f(\boldsymbol{x}) = 0$, t est une raison suffisante de $\boldsymbol{x}$ étant donnée f si et seulement si t est un impliquant premier de $\neg f$. Contrairement aux raisons directes (Izza et al., 2020), les raisons suffisantes ne contiennent aucun attribut redondant. Il est souvent intéressant de se focaliser sur les plus courtes (les raisons suffisantes minimales) car la concision est généralement une propriété souhaitable des explications (selon le principe dit du rasoir d'Occam).

Définition 4 (Raison suffisante minimale)**.** *Soient* $f \in \mathcal{F}_n$ *et* $\boldsymbol{x} \in \{0,1\}^n$*. Une* raison suffisante minimale *de* $\boldsymbol{x}$ *étant donnée* f *est une raison suffisante de* $\boldsymbol{x}$ *étant donnée* f *qui contient un nombre minimal de littéraux.*

Contrairement aux raisons directes et suffisantes (éventuellement minimales), qui visent à expliquer le classement d'une instance $\boldsymbol{x}$ par le classeur f, les explications contrastives sont utiles lorsque l'instance $\boldsymbol{x}$ n'a pas été classée par f comme attendu. On souhaite déterminer des sous-ensembles minimaux des attributs qui, une fois niés dans $\boldsymbol{x}$, sont suffisants pour obtenir une instance positive (resp. négative) selon f si $\boldsymbol{x}$ est une instance négative (resp. positive) selon f.

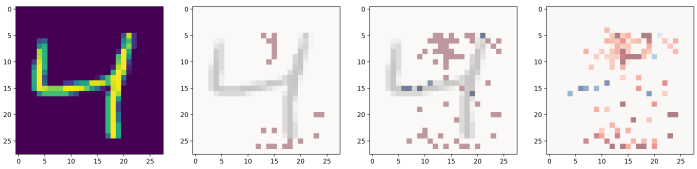

FIG. 2: Une instance `mnist` (à gauche), puis deux raisons suffisantes, dont une minimale (la première des deux) et enfin une carte de chaleur explicative de cette instance (à droite).

Définition 5 (Explication contrastive)**.** *Soient $f \in \mathcal{F}_n$ et $\boldsymbol{x} \in \{0,1\}^n$ telles que $f(\boldsymbol{x}) = 1$ (resp. $f(\boldsymbol{x}) = 0$). Une* explication contrastive *de $\boldsymbol{x}$ étant donnée f est un terme t sur X_n tel que $t \subseteq t_{\boldsymbol{x}}$, $t_{\boldsymbol{x}} \setminus t$ n'est pas un impliquant de f (resp. $\neg f$), et pour tout $\ell \in t$, $t \setminus \{\ell\}$ ne satisfait pas cette dernière condition.*

Exemple 2. *En reprenant l'exemple 1, nous pouvons observer que $T(\boldsymbol{x}) = 1$ pour l'instance $\boldsymbol{x} = (1,1,1,1)$. La raison directe de $\boldsymbol{x}$ est $t^T_{\boldsymbol{x}} = x_1 \wedge x_2 \wedge x_3 \wedge x_4$. $x_1 \wedge x_4$ et $x_2 \wedge x_3 \wedge x_4$ sont deux raisons suffisantes pour $\boldsymbol{x}$. $x_1 \wedge x_4$ est l'unique raison suffisante minimale de $\boldsymbol{x}$. x_4, $x_1 \wedge x_2$, et $x_1 \wedge x_3$ sont des explications contrastives de $\boldsymbol{x}$. En effet, l'instance $(1,1,1,0)$ qui diffère de $\boldsymbol{x}$ sur x_4 seulement n'est pas classée par T comme $\boldsymbol{x}$ l'est ($(1,1,1,0)$ est classée comme une instance négative).*

3 Calculer des explications

Le nombre de raisons suffisantes d'une instance peut être exponentiel. Dans ce qui suit, nous montrons que, même pour la classe restreinte des arbres de décision ayant une profondeur logarithmique, une instance $\boldsymbol{x}$ peut avoir un nombre exponentiel de raisons suffisantes :

Proposition 1. *Il existe un arbre de décision $T \in \mathtt{DT}_n$ de profondeur $\log_2(n+1)$ tel que pour tout $\boldsymbol{x} \in \{0,1\}^n$, le nombre de raisons suffisantes de $\boldsymbol{x}$ étant donné T est supérieur à $\lfloor \frac{3}{2}^{\frac{n+1}{2}} \rfloor$.*

Par définition, le nombre de raisons suffisantes minimales de $\boldsymbol{x}$ ne peut pas être plus grand que celui de ses raisons suffisantes. Mais se limiter aux raisons suffisantes minimales ne permet pas de réduire à coup sûr leur nombre de manière significative car une instance peut avoir un nombre exponentiel de raisons suffisantes minimales :

Proposition 2. *Pour tout $n \in \mathbb{N}$ tel que n est impair, il existe un arbre de décision $T \in \mathtt{DT}_n$ de profondeur $\frac{n+1}{2}$ tel que T contient $2n+1$ nœuds et une instance $\boldsymbol{x} \in \{0,1\}^n$ tels que le nombre de raisons suffisantes minimales de $\boldsymbol{x}$ étant donné T est égal à $2^{\sqrt{n-1}}$.*

En pratique aussi, le nombre de raisons suffisantes peut être très important. Par ailleurs, de telles raisons peuvent ne partager aucun littéral. La figure 2 présente une instance `mnist` (à gauche) qui possède 569,351,040 raisons suffisantes. Certaines de ces raisons suffisantes sont très dissemblables. Pour illustrer cela, nous présentons deux raisons suffisantes, dont une minimale. Elles diffèrent sur de nombreux attributs (les points bleus (resp. rouges) correspondent à des pixels allumés (resp. éteints)).

Pour de telles instances, calculer l'ensemble de toutes les raisons suffisantes n'est pas toujours faisable. Par ailleurs, quand le calcul est possible mais que le nombre de raisons suffisantes est énorme, leur ensemble (interprété de manière disjonctive et appelé raison complète de l'instance (Darwiche et Hirth, 2020)), peut difficilement être considéré comme intelligible. Enfin, vu le nombre de raisons suffisantes et leur diversité, se contenter de calculer l'une d'elles n'est pas toujours informatif. Ainsi, il faut concevoir des approches pour synthétiser l'ensemble de ces raisons en évitant ces deux écueils (calculatoire et informationnel).

Synthétiser l'ensemble des raisons suffisantes. Nous introduisons maintenant les notions d'*attributs nécessaires / non pertinents* pour pallier ces problèmes. Ces notions font écho à celles qui ont été mises en avant dans (Eiter et Gottlob, 1995) pour l'abduction logique.

Définition 6 (Attributs explicatifs)**.** *Soient $f \in \mathcal{F}_n$ et $\boldsymbol{x} \in \{0,1\}^n$ une instance. Soit e un type d'explication (abductif ou contrastif).*

- *Un littéral ℓ sur X_n est un* attribut nécessaire *pour la famille d'explications e de $\boldsymbol{x}$ étant donnée f si et seulement si ℓ appartient à toute explication t de $\boldsymbol{x}$ étant donnée f telle que t est de type e. $Nec_e(\boldsymbol{x}, f)$ désigne l'ensemble de toutes les attributs nécessaires de la famille des explications e de $\boldsymbol{x}$ étant donnée f.*
- *Un littéral ℓ sur X_n est un* attribut pertinent *pour la famille d'explications e de $\boldsymbol{x}$ étant donnée f si et seulement si ℓ appartient à au moins une explication t de $\boldsymbol{x}$ étant donnée f telle que t est de type e. $Rel_e(\boldsymbol{x}, f)$ désigne l'ensemble de tous les attributs pertinents de la famille des explications e de $\boldsymbol{x}$ étant donnée f. $Irr_e(\boldsymbol{x}, f)$, qui est le complément de $Rel_e(\boldsymbol{x}, f)$ dans l'ensemble de tous les littéraux sur X_n, désigne l'ensemble des tous les attributs non pertinents pour la famille d'explications e de $\boldsymbol{x}$ étant donnée f.*

Les attributs nécessaires (resp. non pertinents) pour la famille s des raisons suffisantes de $\boldsymbol{x}$ étant donnée f sont les attributs les plus (resp. les moins) importants pour expliquer le classement de $\boldsymbol{x}$ par f, puisqu'ils appartiennent à toute (resp. aucune) raison suffisante de $\boldsymbol{x}$.

Pour une raison suffisante t de $\boldsymbol{x}$ étant donnée f, lorsque le cardinal de t privé des attributs de $Nec_s(\boldsymbol{x}, f)$ est petit, et que le cardinal de la différence symétrique entre t et $Rel_s(\boldsymbol{x}, f)$ est également petit, t peut être considérée comme une bonne représentation de la raison complète de $\boldsymbol{x}$ étant donnée f puisqu'une raison suffisante t' de $\boldsymbol{x}$ étant donnée f qui diffère beaucoup de t ne peut pas exister.

Quel que soit le nombre de raisons suffisantes d'une instance $\boldsymbol{x}$ étant donné T, $Nec_s(\boldsymbol{x}, f)$, $Rel_s(\boldsymbol{x}, f)$, et $Irr_s(\boldsymbol{x}, f)$ peuvent être calculés efficacement :

Proposition 3. *Soient $T \in \mathtt{DT}_n$ et $\boldsymbol{x} \in \{0,1\}^n$. Calculer $Nec_s(\boldsymbol{x}, T)$, $Rel_s(\boldsymbol{x}, f)$, et $Irr_s(\boldsymbol{x}, T)$ peut se faire en temps $\mathcal{O}((n + |T|) \times |T|)$.*

Définition 7 (Pouvoir explicatif)**.** *Soient $f \in \mathcal{F}_n$, et $\boldsymbol{x} \in \{0,1\}^n$ une instance. Soit e un type d'explication (abductif ou contrastif), et $E_e(\boldsymbol{x}, f)$ l'ensemble de tous les explications de $\boldsymbol{x}$ étant donnée f qui sont de type e. Le* pouvoir explicatif *selon e d'un littéral ℓ sur X_n pour $\boldsymbol{x}$ étant donnée f est donné par*

$$Imp_e(\ell, \boldsymbol{x}, f) = \frac{\#(\{t \in E_e(\boldsymbol{x}, f) : \ell \in t\})}{\#(E_e(\boldsymbol{x}, f))}.$$

Exemple 3. *Revenons à l'exemple 1 : nous avons* $Nec_s(\boldsymbol{x}, T) = \{x_4\}$*, et* $Rel_s(\boldsymbol{x}, T) = \{x_1, x_2, x_3, x_4\}$*. Nous avons aussi* $Imp_s(x_4, \boldsymbol{x}, T) = 1$*,* $Imp_s(x_1, \boldsymbol{x}, T) = Imp_s(x_2, \boldsymbol{x}, T) = Imp_s(x_3, \boldsymbol{x}, T) = \frac{1}{2}$*, et* $Imp_s(\ell, \boldsymbol{x}, T) = 0$ *pour tout littéral* $\ell \in \{\overline{x_1}, \overline{x_2}, \overline{x_3}, \overline{x_4}\}$.

La notion de pouvoir explicatif ne doit pas être confondue avec la notion d'importance des attributs [1] (qui peut être définie et évaluée de différentes manières) : la première est locale (c-à-d relative à une instance) et non globale, elle concerne des littéraux et non des variables, et il s'agit d'évaluer l'importance du littéral pour la tâche d'explication, pas pour celle de prédiction.

Une approche pour calculer $Imp_s(\ell, \boldsymbol{x}, T)$**.** On sait que $sr(\boldsymbol{x}, T)$ est par construction un ensemble d'impliquants premiers de $g = \{c \cap t_{\boldsymbol{x}} : c \in \texttt{CNF}(T)\}$. Nous exploitons la réduction présentée dans (Jabbour et al., 2014) montrant comment associer en temps polynomial une formule CNF (ici, g) à une autre formule h (sur un ensemble distinct de variables), tel que les modèles de h correspondent bijectivement aux impliquants premiers de g.

On utilise alors le compteur de modèles à base de compilation D4 (Lagniez et Marquis, 2017) pour compiler g dans le langage d−DNNF (Darwiche, 2001), nous permettant de calculer ensuite en temps polynomial en la taille de g à la fois le nombre de raisons suffisantes et le pouvoir explicatif de chaque littéral. En effet, le langage d−DNNF supporte la requête de comptage de modèle et la transformation de conditionnement (Darwiche et Marquis, 2002).

Lorsque $Imp_e(\ell, \boldsymbol{x}, T)$ a été calculé pour chaque ℓ, on peut facilement générer des cartes de chaleur explicatives. La figure 2 la plus à droite montre une telle carte pour une instance mnist avec 12 littéraux nécessaires et 105 littéraux pertinents. Les pixels bleus (resp. rouges) correspondent aux littéraux positifs (resp. négatifs) dans l'instance, et l'intensité de la couleur vise à refléter le pouvoir explicatif du littéral correspondant.

Énumérer les raisons suffisantes minimales. Afin de synthétiser l'ensemble des raisons suffisantes, nous pouvons nous focaliser sur les raisons suffisantes minimales. En effet, bien que l'ensemble des raisons suffisantes minimales d'une instance étant donné un arbre de décision puisse être exponentiellement grand, le nombre de raisons suffisantes minimales ne peut pas dépasser le nombre de raisons suffisantes, et, en pratique, il peut être nettement inférieur. Toutefois, contrairement aux raisons suffisantes qui peuvent être générées en temps polynomial (Izza et al., 2020), calculer les raisons minimales n'est pas facile :

Proposition 4. *Soient* $T \in \texttt{DT}_n$ *et* $\boldsymbol{x} \in \{0,1\}^n$*. Calculer une raison suffisante minimale de* $\boldsymbol{x}$ *étant donné* T *est* NP*-difficile.*

Malgré ce résultat d'intraitabilité dans le cas général, calculer un ensemble de raisons suffisantes minimales est possible dans de nombreux cas pratiques. Pour cela, on s'appuie sur les progrès récents réalisés en optimisation combinatoire autour de SAT. Rappelons tout d'abord qu'un problème d'optimisation PARTIAL MAXSAT consiste en une paire $(C_{\text{soft}}, C_{\text{hard}})$ où C_{soft} et C_{hard} sont des ensembles (finis) de clauses. L'objectif est de déterminer quand elle existe une affectation des variables qui maximise le nombre de clauses satisfaites de C_{soft}, tout en satisfaisant toutes les clauses de C_{hard}. On peut utiliser un solveur PARTIAL MAXSAT pour calculer des raisons suffisantes minimales :

1. Dans cet article, nous utilisons le mot attribut dans deux contextes : l'attribut en tant que descripteur de l'instance (ici, une variable propositionnelle) et le couple (attribut,valeur), représenté ici par un littéral.

Proposition 5. *Soient T un arbre de décision dans $\mathtt{DT}_n$ et $\boldsymbol{x} \in \{0,1\}^n$ une instance tels que $T(\boldsymbol{x}) = 1$. Soit $(C_{\text{soft}}, C_{\text{hard}})$ une instance du problème* PARTIAL MAXSAT *telle que :*

$$C_{\text{soft}} = \{\overline{x_i} : x_i \in t_{\boldsymbol{x}}\} \cup \{x_i : \overline{x}_i \in t_{\boldsymbol{x}}\} \text{ et } C_{\text{hard}} = \{c \cap t_{\boldsymbol{x}} : c \in \mathtt{CNF}(T)\}.$$

L'intersection de $t_{\boldsymbol{x}}$ avec $t_{\boldsymbol{x}^*}$, où $\boldsymbol{x}^*$ est une solution optimale de $(C_{\text{hard}}, C_{\text{soft}})$, est une raison suffisante minimale de $\boldsymbol{x}$ étant donné T.

On peut également exploiter un solveur PARTIAL MAXSAT pour calculer un nombre prédéfini de raisons suffisantes minimales. On génère pour cela une première raison t, puis on ajoute à C_{hard} la négation de t ainsi qu'une contrainte de cardinalité pour s'assurer que les prochaines raisons calculées auront la même taille que t, et on réitère ce processus jusqu'à ce que le nombre de raisons espéré soit atteint ou que plus aucune solution n'existe.

Calculer toutes les explications contrastives Il a été montré que les raisons suffisantes et les explications contrastives sont liées par dualité (Ignatiev et al., 2020). On peut exploiter cette dualité pour passer d'un type d'explication à l'autre en utilisant des algorithmes pour calculer des ensembles intersectants minimaux (*minimal hitting sets*) (Reiter, 1987).Cependant, dans le cas des arbres de décision, une approche beaucoup plus efficace pour dériver toutes les explications contrastives peut être mise en œuvre. En effet, on peut calculer l'ensemble les explications contrastives d'une instance en temps polynomial (voir aussi **?**).

Proposition 6. *L'ensemble de toutes les explications contrastives de $\boldsymbol{x} \in \{0,1\}^n$ étant donné un arbre de décision $T \in \mathtt{DT}_n$ peut être calculé en temps polynomial en $n + |T|$ comme $min(\{c \cap t_{\boldsymbol{x}} : c \in \mathtt{CNF}(f)\}, \subseteq)$.*

Exemple 4. *Continuons avec l'exemple 1. On a $\mathtt{CNF}(T) = \{x_1 \vee x_2, x_1 \vee \overline{x_2} \vee x_3, x_1 \vee \overline{x_2} \vee \overline{x_3} \vee x_4, \overline{x_1} \vee x_2 \vee x_3 \vee x_4, \overline{x_1} \vee x_2 \vee \overline{x_3} \vee x_4, \overline{x_1} \vee \overline{x_2} \vee x_3 \vee x_4, \overline{x_1} \vee \overline{x_2} \vee \overline{x_3} \vee x_4\}$. Ainsi, avec $\boldsymbol{x} = (1,1,1,1)$, on a $min(\{c \cap t_{\boldsymbol{x}} : c \in \mathtt{CNF}(f)\}, \subseteq) = \{x_1 \vee x_2, x_1 \vee x_3, x_4\}$, qui correspond à l'ensemble des explications contrastives $x_1 \wedge x_2$, $x_1 \wedge x_3$, x_4 de $\boldsymbol{x}$ étant donné T.*

4 Expérimentations

Protocole expérimental. Nous avons considéré 90 jeux de données (aussi appelés datasets ou benchmarks dans la suite) bien connus et disponibles sur Kaggle (`www.kaggle.com`), OpenML (`www.openml.org`), et UCI (`archive.ics.uci.edu/ml/`). `mnist38` et `mnist49` sont deux sous-ensembles de `mnist`, restreints aux instances des chiffres 3 et 8 (resp. 4 et 9). Certains datasets concernent une problématique de classification multi-classe, nous avons alors utilisé la politique « un contre tous »pour les traiter : toutes les classes sauf celle visée sont considérées comme la classe complémentaire de la cible. Pour les attributs numériques, aucun pré-traitement des données n'a été réalisé : les attributs ont été binarisés en ligne par l'algorithme d'apprentissage d'arbre de décision utilisé. Pour chaque benchmark b, nous avons effectué une validation croisée à 10 blocs, de manière classique. La performance de la classification pour b est évaluée comme la performance moyenne obtenue sur les 10 arbres de décisions générés. L'algorithme CART a été utilisé pour apprendre les arbres de décision

TAB. 1: Résultats empiriques sur 12 datasets.

	Decision Tree			|Sufficient|		|Minimal|		#Nec. Features		#Rel. Features	
Dataset	%A	#N	#B	med	max	med	max	med	max	med	max
recidivism	63.35	13808.60	148.30	14	23	14	23	6	20	144	168
adult	81.46	12900.20	2982.40	15	45	15	45	6	27	2248	2407
bank marketing	87.42	5984.20	1333.40	13	20	12	20	3	11	1036	1064
bank	89.00	5517.60	985.10	13	22	13	22	4	17	794	856
lending loan	73.77	2608.80	1128.60	16	34	16	34	8	20	848	891
contraceptive	48.47	1247.20	87.50	11	21	11	21	8	20	88	99
compas	65.84	1225.60	46.00	6	15	6	15	3	13	44	56
christine	62.93	848.40	423.50	13	44	13	44	8	38	296	327
farm-ads	87.56	542.60	264.40	20	101	20	101	15	92	201	223
mnist49	95.40	554.00	275.90	23	32	23	32	10	23	205	240
spambase	91.93	527.00	259.60	15	31	15	31	9	27	194	230
mnist38	96.10	499.60	248.40	18	28	18	28	7	23	185	215

	#Sufficient		#Contrastive		|Contrastive|		#Minimal	
Dataset	med	max	med	max	med	max	med	max
recidivism	-	$\geq$9147225	2218	4732	7	20	2	80
adult	-	$\geq$1101171049856	3045	3438	7	30	3	384
bank marketing	-	$\geq$87859325304	1390	1427	6	19	4	108
bank	-	$\geq$63315924	1282	1506	6	18	4	768
lending loan	-	$\geq$52055382208395454648911721922256	636	685	5	21	3	432
contraceptive	399	26478	407	451	5	15	2	96
compas	37	1172	277	349	5	13	2	38
christine	361024452	2840555408902862336	211	221	4	12	2	768
farm-ads	416416	1547127889920	159	184	3	26	-	$\geq$ 10000
mnist49	2827347768	3048749438423040	137	153	4	17	-	$\geq$ 10000
spambase	396990	9931021676906400	138	157	4	19	4	2304
mnist38	-	$\geq$ 297394157793600	121	141	4	14	32	2048

via son implémentation dans Scikit-Learn. Tous les hyper-paramètres ont été laissés à leurs valeurs par défaut.

Pour chaque benchmark b, chaque arbre de décision T_b, et un sous-ensemble d'au plus 100 instances $\boldsymbol{x}$ prises au hasard dans l'ensemble de test suivant une distribution uniforme, nous avons calculé la taille des raisons suffisantes (en utilisant l'algorithme glouton standard exécuté sur la raison directe $t_{\boldsymbol{x}}^{T_b}$), celle des raisons suffisantes minimales (via l'appel au solveur PARTIAL MAXSAT RC2 (Ignatiev et al., 2018)). Nous avons également calculé le nombre d'attributs nécessaires et pertinents de chaque instance considérée ainsi que le nombre de ses raisons suffisantes (pour ce faire, nous avons utilisé le compteur de modèles `D4` (Lagniez et Marquis, 2017)). Enfin, nous avons calculé le nombre et la taille des explications contrastives ainsi que le nombre de raisons suffisantes minimales (limité à 10000). A partir de là, nous avons extrait des statistiques (médiane, maximum) des différents résultats obtenus. Pour chaque calcul réalisé, nous avons mesuré le temps de calcul correspondant car cela est fondamental pour déterminer dans quelle mesure nos algorithmes sont pratiques. Toutes les expérimentations ont été effectuées sur un ordinateur équipé d'un processeur Intel(R) XEON E5-2637 CPU @ 3.5 GHz and 128 Go de mémoire. Un temps limite (TO) de 100s par instance a été utilisé pour `D4`.

Résultats. Le tableau 1 (haut) donne une synthèse de nos résultats, sur 12 datasets (les datasets sélectionnés sont ceux qui contiennent de nombreuses instances et/ou de nombreux attributs). La colonne la plus à gauche donne le nom du dataset b. Les colonnes $\%A$, $\%N$, et $\#B$ donnent, respectivement, la précision moyenne sur les 10 arbres de décision, le nombre moyen de nœuds dans ces arbres, et le nombre moyen d'attributs binaires. La colonne suivante donne des statistiques (médiane, maximum) sur, respectivement, la taille des raisons suffisantes (|Sufficient|) et de raisons suffisantes minimales (|Minimal|) qui ont été calculées, ainsi que sur

le nombre d'attributs nécessaires (#Nec. Features) et pertinents (#Rel. Features) qui apparaissent dans l'ensemble complet des raisons suffisantes de l'instance. Le tableau 1 (bas) donne la médiane et le maximum (respectivement), du nombre de raisons suffisantes, (#Sufficient), du nombre d'explications contrastives (#Contrastive) et de leurs tailles (|Contrastive|), et finalement du nombre de raisons suffisantes minimales (#Minimal).

En ce qui concerne les temps de calcul, les approches décrites dans les sections précédentes se sont montrées très efficaces en pratique. Ce n'est pas surprenant pour celles reposant sur des algorithmes ayant une complexité polynomiale dans le pire des cas (l'algorithme glouton pour calculer une raison suffisante, celui pour dériver des attributs explicatifs, et celui pour calculer toutes les explications contrastives). C'était moins évident à première vue pour les algorithmes utilisés pour compter le nombre de raisons suffisantes et pour calculer le pouvoir explicatif des attributs. Toutefois, tous les calculs qui ont été lancés ont pris fin en temps voulu pour les 90 datasets, à l'exception de 8 : `bank`, `adult`, `bank_marketing`, `lending loan`, `recidivism`, `mnist38`, `gina`, et `german`. Pour ces datasets, le temps limite de 100s a été atteint pour, respectivement, 993, 932, 241, 130, 47 et 1 (pour les trois datasets restants) instances sur les 1000 considérées. Dans ce cas, la médiane pour les raisons suffisantes n'a pas été calculée. Calculer le nombre de raisons suffisantes pour des instances de `adult` et `bank` semble inaccessible la plupart du temps. Hormis pour ces datasets, le temps médian requis pour calculer le nombre de raisons suffisantes et calculer le pouvoir explicatif des attributs dépasse rarement $1s$. Calculer toutes les raisons suffisantes minimales semblait a priori difficile. Néanmoins, notre approche d'énumération a réussi à dériver *toutes les raisons suffisantes minimales* pour chaque dataset sauf pour `gisette`, `mnist49`, `farm-ads`, et `yeast` . Pour ces datasets, la limite de 10 000 raisons a été atteinte pour, respectivement, 61, 4, 2 et 2 instances (sur 1000). De plus, le temps médian nécessaire pour dériver toutes les raisons suffisantes minimales n'a dépassé $5s$ que pour 2 datasets (`adult` et `gisette`).

Nos expérimentations ont montré que le nombre de raisons suffisantes peut être énorme. Elles ont également mis en évidence que l'algorithme glouton pour dériver une raison suffisante calcule une raison dont la taille est proche de la taille d'une raison suffisante minimale. Nos expérimentations ont aussi montré que le nombre d'attributs pertinents pour une instance est généralement bien inférieur au nombre d'attributs binaires utilisées pour la décrire, et que le nombre d'attributs nécessaires est également nettement inférieur au nombre d'attributs pertinents. Cet écart explique le nombre très élevé de raisons suffisantes. Une différence considérable entre le nombre de raisons suffisantes et le nombre de raisons suffisantes minimales peut également être observé. Enfin, le nombre d'explications contrastives est souvent faible, ce qui est une bonne nouvelle du point de vue de l'intelligibilité.

5 Conclusion

À la lumière de nos résultats, il est clair que le pouvoir explicatif des arbres de décision va bien au-delà de leur capacité à générer efficacement des raisons directes. Pour un arbre de décision, on peut calculer efficacement (dans la plupart des cas) le pouvoir explicatif des attributs et les raisons suffisantes minimales pour une instance. Pour un arbre de décision, aborder la question « pourquoi pas ? » apparaît aussi comme plus facile qu'aborder à la question « pourquoi ? ». Calculer l'ensemble de toutes les raisons suffisantes d'une instance est généralement hors de portée, tandis que calculer l'ensemble de toutes les raisons contrastives est traitable. Ainsi, le

langage des arbres de décision apparaît non seulement comme attrayant pour l'apprentissage, mais aussi comme un bon langage cible lorsqu'il faut raisonner sur les diverses formes d'explications (abductives et contrastives) associées aux prédictions faites. Ceci est cohérent avec les résultats présentés dans (Audemard et al., 2020), montrant que d'autres requêtes d'explication et de vérification sont traitables pour les classeurs de type arbre de décision.

Remerciements

Merci aux relecteurs pour leurs commentaires. Ils nous ont été précieux pour améliorer l'article. Le travail correspondant a été réalisé dans le cadre de la chaire ANR d'enseignement et de recherche EXPEKCTATION (ANR-19-CHIA-0005-01).

Références

Audemard, G., F. Koriche, et P. Marquis (2020). On tractable XAI queries based on compiled representations. In *Proc. of KR'20*, pp. 838–849.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Breiman, L., J. H. Friedman, R. A. Olshen, et C. J. Stone (1984). *Classification and Regression Trees*. Wadsworth.

Breiman, L. et N. Shang (1996). Born again trees. Technical report, https ://www.stat.berkeley.edu/ breiman/BAtrees.pdf.

Darwiche, A. (2001). Decomposable negation normal form. *Journal of the Association for Computing Machinery 48*(4), 608–647.

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

Darwiche, A. et P. Marquis (2002). A knowledge compilation map. *Journal of Artificial Intelligence Research 17*, 229–264.

Eiter, T. et G. Gottlob (1995). The complexity of logic-based abduction. *Journal of the Association for Computing Machinery 42*(1), 3–42.

Frosst, N. et G. E. Hinton (2017). Distilling a neural network into a soft decision tree. In *Proc. of the First International Workshop on Comprehensibility and Explanation in AI and ML*, Volume 2071 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, et D. Pedreschi (2019). A survey of methods for explaining black box models. *ACM Computing Surveys 51*(5), 93 :1–93 :42.

Hooker, S., D. Erhan, P.-J. Kindermans, et B. Kim (2019). A benchmark for interpretability methods in deep neural networks. In *Proc. of NeurIPS'19*, pp. 9737–9748.

Huysmans, J., K. Dejaeger, C. Mues, J. Vanthienen, et B. Baesens (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decis. Support Syst. 51*(1), 141–154.

Ignatiev, A., A. Morgado, et J. Marques-Silva (2018). PySAT : A Python toolkit for prototyping with SAT oracles. In *Proc. of SAT'18*, pp. 428–437.

Ignatiev, A., N. Narodytska, N. Asher, et J. Marques-Silva (2020). On relating 'why ?' and 'why not ?' explanations. *CoRR abs/2012.11067*.

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., A. Ignatiev, et J. Marques-Silva (2020). On explaining decision trees. *CoRR abs/2010.11034*.

Jabbour, S., J. Marques-Silva, L. Sais, et Y. Salhi (2014). Enumerating prime implicants of propositional formulae in conjunctive normal form. In *Proc. of JELIA'14*, pp. 152–165.

Lagniez, J.-M. et P. Marquis (2017). An Improved Decision-DNNF Compiler. In *Proc. of IJCAI'17*, pp. 667–673.

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

Lundberg, S. et S.-I. Lee (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et R. Garnett (Eds.), *Proc. of NIPS'17*, pp. 4765–4774.

Miller, T. (2019). Explanation in artificial intelligence : Insights from the social sciences. *Artificial Intelligence 267*, 1–38.

Molnar, C. (2019). *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable*. Leanpub.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning 1*(1), 81–106.

Reiter, R. (1987). A theory of diagnosis from first principles. *Artificial Intelligence 32*, 57–95.

Ribeiro, M., S. Singh, et C. Guestrin (2016). "Why should I trust you ?" : Explaining the predictions of any classifier. In *Proc. of KDD'16*, pp. 97–101.

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

Shih, A., A. Darwiche, et A. Choi (2019). Verifying binarized neural networks by Angluin-style learning. In *Proc. of SAT'19*, pp. 354–370.

Summary

Decision trees are a learning model suitable for applications where the interpretability of decisions is of paramount importance. Here we examine the ability of binary decision trees to extract, minimize, and count abductive explanations and contrastive explanations. We prove that the set of all irredundant abductive explanations (alias sufficient reasons) of an instance can be of exponential size Therefore, generating the full set may turn out to be out of reach. Moreover, two sufficient reasons of the same instance can differ on all their attributes. Thus, the computation of a single sufficient reason only gives a fragmentary view of the possible explanations. We present the notions of necessary / relevant attribute for an explanation and the notion of explanatory importance of an attribute and we show that these notions can be useful to derive a synthetic view of the set of all sufficient reasons of an instance.