Trading Complexity for Sparsity in Random Forest Explanations

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

Trading Complexity for Sparsity in Random Forest Explanations

Gilles Audemard[1], **Steve Bellart**[1], **Louènas Bounia**[1],
Frédéric Koriche[1], **Jean-Marie Lagniez**[1], **Pierre Marquis**[1,2]

[1]Univ. Artois, CNRS, CRIL, F-62300 Lens, France
[2]Institut Universitaire de France
{audemard, bellart, bounia, koriche, lagniez, marquis}@cril.fr

Abstract

Random forests have long been considered as powerful model ensembles in machine learning. By training multiple decision trees, whose diversity is fostered through data and feature subsampling, the resulting random forest can lead to more stable and reliable predictions than a single decision tree. This however comes at the cost of decreased interpretability: while decision trees are often easily interpretable, the predictions made by random forests are much more difficult to understand, as they involve a majority vote over multiple decision trees. In this paper, we examine different types of *reasons* that explain "why" an input instance is classified as positive or negative by a Boolean random forest. Notably, as an alternative to *prime-implicant explanations* taking the form of subset-minimal implicants of the random forest, we introduce *majoritary reasons* which are subset-minimal implicants of a strict majority of decision trees. For these abductive explanations, the tractability of the generation problem (finding one reason) and the optimization problem (finding one minimum-sized reason) are investigated. Unlike prime-implicant explanations, majoritary reasons may contain redundant features. However, in practice, prime-implicant explanations - for which the identification problem is DP-complete - are slightly larger than majoritary reasons that can be generated using a simple linear-time greedy algorithm. They are also significantly larger than *minimum-sized* majoritary reasons which can be approached using an anytime PARTIAL MAXSAT algorithm.

Introduction

Over the past two decades, rapid progress in statistical machine learning has led to the deployment of models endowed with remarkable predictive capabilities. Yet, as the spectrum of applications using statistical learning models becomes increasingly large, explanations for why a model is making certain predictions are ever more critical. For example, in medical diagnosis, if some model predicts that an image is malignant, then the physician may need to know which features in the image have led to this classification. Similarly, in the banking sector, if some model predicts that a customer commits fraud, then the banker might want to know why. Therefore, deriving explanations for why certain predictions are made is essential for securing user confidence in machine learning technologies (Miller 2019; Molnar 2019).

This paper focuses on classifications made by *random forests*, a popular ensemble learning method that constructs multiple randomized decision trees during the training phase, and predicts by taking a majority vote over the base classifiers (Breiman 2001). Since decision tree randomization is achieved by essentially coupling data subsampling (or bagging) and feature subsampling, random forests are easy to implement, with few tuning parameters. Furthermore, they often make accurate and robust predictions in practice, even for small data samples and high-dimensional feature spaces (Biau 2012). For these reasons, random forests have been used in various applications including, among others, computer vision (Criminisi and Shotton 2013), crime prediction (Bogomolov et al. 2014), ecology (Cutler et al. 2007), genomics (Chen and Ishwaran 2012), and medical diagnosis (Azar et al. 2014).

However, random forests are often considered far less interpretable than decision trees. Thus, while many XAI queries are tractable for decision trees, they are not tractable for random forests (Audemard et al. 2021). Especially, the prediction made on a given data instance can be easily interpreted by reading the *direct reason* furnished by the classifier. For a decision tree, it is the unique root-to-leaf path that covers the instance (also known as its path-restricted explanation (Izza, Ignatiev, and Marques-Silva 2020)). By contrast, no such direct reason has been defined so far for a random forest. More generally, a key issue in random forests is to infer *abductive explanations*, that is, to capture in concise terms why a data instance is classified as positive or negative by the model ensemble.

Related Work. Explaining random forest predictions has received increasing attention in recent years (Bénard et al. 2021; Choi et al. 2020; Izza and Marques-Silva 2021). In the classification setting, Choi et al. (2020) and Izza & Marques-Silva (2021) have focused on *prime-implicant explanations*, also known as *sufficient reasons* (Darwiche and Hirth 2020). Informally, if we view any random forest classifier as a Boolean function f, then a prime-implicant explanation for classifying a data instance $\boldsymbol{x}$ as positive by f is a subset-minimal implicant t of f covering $\boldsymbol{x}$. By construction, t is an abductive explanation involving only relevant features,

since removing any feature from t would question the fact that t explains the way $\boldsymbol{x}$ is classified by f. Note that if f is described by a single decision tree, then generating a prime-implicant explanation for any data instance $\boldsymbol{x}$ can be done in linear time. Yet, in the general case where f is represented by an arbitrary number of decision trees, the problem of determining whether a given term is a prime-implicant explanation for an instance given a random forest has recently been shown DP-complete (Izza and Marques-Silva 2021). Despite this intractability statement, the empirical results reported by the authors show that a MUS-based algorithm for computing such explanations can prove efficient in practice.

In addition to *model-based* explanations described above, *model-agnostic* explanations can be applied to random forests. Notably, the LIME method (Ribeiro, Singh, and Guestrin 2016) extrapolates a linear threshold function g from the behavior of the random forest f around an input instance $\boldsymbol{x}$. For the ANCHOR method (Ribeiro, Singh, and Guestrin 2018), the extrapolated model g takes the form of a decision rule. Yet, even if in both cases a prime implicant of g can be easily computed, the resulting explanation is *not* guaranteed to be an abductive explanation for $\boldsymbol{x}$ given f since g is an approximation of f.

Contributions. In this paper, we introduce new notions of abductive explanations: *direct reasons*, which extend to the case of random forests the corresponding notion defined for decision trees, and *majoritary reasons*, which are abductive explanations taking into account the averaging rule of random forests. Informally, a majoritary reason for classifying an instance $\boldsymbol{x}$ as positive by some random forest F is a term t that covers $\boldsymbol{x}$ and is a subset-minimal implicant of a strict majority of decision trees in F.

What makes direct and majoritary reasons valuable is the possibility of inferring them in a tractable way, whilst there is no similar tractability result when dealing with sufficient reasons, unless P = NP. In the following, we examine the tractability of both the generation problem (finding one explanation) and the optimization problem (finding one minimum-sized explanation) for direct and majoritary reasons. Direct reasons (that coincide with minimum-sized direct reasons) and majoritary reasons can be derived in polynomial time. By contrast, identifying minimum-sized majoritary reasons is NP-complete, and identifying minimum-sized prime-implicant explanations is Σ_2^p-complete.

Based on these results, we provide algorithms for deriving random forest explanations, enabling an empirical comparison. Our experiments made on standard benchmarks assess both the runtime complexity of finding abductive explanations and the sparsity of such explanations (i.e., how much parsimonious they are). In a nutshell, majoritary reasons and minimum-sized majoritary reasons offer interesting compromises in comparison to, respectively, prime-implicant explanations and minimum-sized prime-implicant explanations. Indeed, even if in theory for every majoritary reason, there exists a prime-implicant explanation that is not larger, in practice the size of majoritary reasons are generally smaller than those of prime-implicant explanations. Furthermore, the computational effort to be spent for deriving majoritary reasons is smaller than the one required by prime-implicant explanations. Similarly, in practice, minimum-sized majoritary reasons outperform minimum-sized prime-implicant explanations, in the sense that deriving minimum-sized prime-implicant explanations is often too computationally demanding. Using an *anytime* PARTIAL MAXSAT solver for minimizing the size of majoritary reasons, we show how to derive abductive explanations which are typically shorter than all other forms of abductive explanations considered in the paper. A full-proof version of the paper is available at www.cril.univ-artois.fr/expekctation/papers.html.

Preliminaries

For an integer n, let $[n] = \{1, \cdots, n\}$. By $\mathcal{F}_n$ we denote the class of all Boolean functions from $\{0,1\}^n$ to $\{0,1\}$, and we use $X_n = \{x_1, \cdots, x_n\}$ to denote the set of input Boolean variables. Any Boolean vector $\boldsymbol{x} \in \{0,1\}^n$ is called an *instance*. For any function $f \in \mathcal{F}_n$, an instance $\boldsymbol{x} \in \{0,1\}^n$ is called a *positive instance* of f if $f(\boldsymbol{x}) = 1$, and a *negative instance* if $f(\boldsymbol{x}) = 0$.

We refer to f as a propositional formula when it is described using the Boolean connectives $\wedge$ (conjunction), $\vee$ (disjunction) and $\neg$ (negation), together with the constants 1 (true) and 0 (false). f is *satisfiable* if it has a positive instance, and it is *unsatisfiable* otherwise. A *literal* l_i is a variable x_i or its negation $\neg x_i$, also denoted $\overline{x}_i$. A *term* t is a conjunction of literals, and a *clause* c is a disjunction of literals. A DNF *formula* is a disjunction of terms and a CNF *formula* is a conjunction of clauses. The set of variables occurring in a formula f is denoted $Var(f)$. In the following, we shall often treat instances as terms, and terms as sets of literals. For an assignment $\boldsymbol{z} \in \{0,1\}^n$, the corresponding term is

$$t_{\boldsymbol{z}} = \bigwedge_{i=1}^{n} x_i^{z_i} \text{ where } x_i^0 = \overline{x}_i \text{ and } x_i^1 = x_i$$

A term t *covers* an assignment $\boldsymbol{z}$ if $t \subseteq t_{\boldsymbol{z}}$. An *implicant* of a Boolean function f is a term that implies f, that is, a term t such that $f(\boldsymbol{z}) = 1$ for every assignment $\boldsymbol{z}$ covered by t. A *prime implicant* of f is an implicant t of f such that no proper subset of t is an implicant of f.

A (Boolean) *decision tree* on X_n is a binary tree T, each of whose internal nodes is labeled with one of n input variables, and whose leaves are labeled 0 or 1. Without loss of generality, every variable is supposed to occur at most once on any root-to-leaf path. The value $T(\boldsymbol{x})$ of T on an input instance $\boldsymbol{x}$ is given by the label of the leaf reached from the root as follows: at each node go to the left (resp. right) child if the input value of the corresponding variable is 0 (resp. 1). A (Boolean) *random forest* on X_n is an ensemble $F = \{T_1, \cdots, T_m\}$, where each T_i ($i \in [m]$) is a decision tree on X_n, and such that the value $F(\boldsymbol{x})$ is given by

$$F(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \frac{1}{m}\sum_{i=1}^{m} T_i(\boldsymbol{x}) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

The *size* of F is given by $|F| = \sum_{i=1}^{m} |T_i|$, where $|T_i|$ is the number of nodes occurring in T_i. The class of decision trees

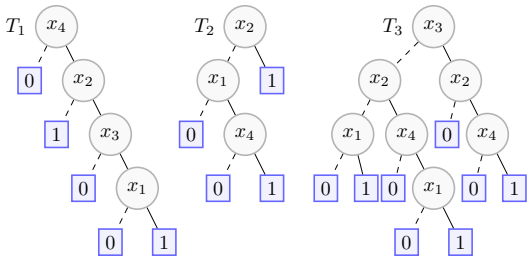

Figure 1: A random forest for recognizing *Cattleya* orchids. The left (resp. right) child of any decision node labelled by x_i corresponds to the assignment of x_i to 0 (resp. 1).

on X_n is denoted $\mathtt{DT}_n$, and the class of random forests with at most m decision trees (with $m \geq 1$) over $\mathtt{DT}_n$ is denoted $\mathtt{RF}_{n,m}$. Finally, $\mathtt{RF}_n = \bigcup_{m \geq 1} \mathtt{RF}_{n,m}$ and $\mathtt{RF} = \bigcup_{n \geq 1} \mathtt{RF}_n$.

Example 1. *The random forest $F = \{T_1, T_2, T_3\}$ in Figure 1 is composed of three decision trees. It separates Cattleya orchids from other orchids using the following features: x_1: "has fragrant flowers", x_2: "has one or two leaves", x_3: "has large flowers", and x_4: "is sympodial".*

We conclude this section with two important properties of random forests, that are useful to prove forthcoming results. The first property is related to the fact that any decision tree T can be transformed into its negation $\neg T \in \mathtt{DT}_n$, by simply inverting the labels of its leaves. Negating a random forest can also be achieved in polynomial time:

Proposition 1. *There exists a linear-time algorithm that computes a random forest $\neg F \in \mathtt{RF}_{n,m}$ equivalent to the negation of a given random forest $F \in \mathtt{RF}_{n,m}$.*

Furthermore, it is well-known that any decision tree T can be encoded in linear time into an equivalent disjunction of terms $\mathtt{DNF}(T)$, where each term coincides with a 1-path (i.e., a path from the root to a leaf labeled with 1), but also as a conjunction of clauses $\mathtt{CNF}(T)$, where each clause is the negation of a term describing a 0-path. Yet, when switching to random forests, the picture is quite different:

Proposition 2. *Any* CNF *or* DNF *formula can be converted in linear time into an equivalent random forest, but there is no polynomial-space translation from* RF *to* CNF *or to* DNF.

Random Forest Explanations

The key focus of this study is to explain *why* a random forest classifies some data instance as positive or negative. This calls for a notion of abductive explanation.[1] Specifically, an *abductive explanation* for an instance $\boldsymbol{x} \in \{0,1\}^n$ given a Boolean function $f \in \mathcal{F}_n$ is a term t that covers $\boldsymbol{x}$ and is an implicant of f (resp. $\neg f$) if $f(\boldsymbol{x}) = 1$ (resp. $f(\boldsymbol{x}) = 0$). Such an abductive explanation always exists, since $t = t_{\boldsymbol{x}}$ is such a (trivial) explanation. So, in the rest of this section, we shall mainly concentrate on *sparse* forms of abductive explanations.

[1] Unlike (Ignatiev, Narodytska, and Marques-Silva 2019), we do not require those explanations to be minimal with respect to set inclusion, in order to keep the concept distinct (and actually more general) than the one of prime-implicant explanations.

Direct Reasons. For a decision tree $T \in \mathtt{DT}_n$ and a data instance $\boldsymbol{x} \in \{0,1\}^n$, the *direct reason* for $\boldsymbol{x}$ given T is the term $t^T_{\boldsymbol{x}}$ corresponding to the unique root-to-leaf path of T that covers $\boldsymbol{x}$. This simple form of abductive explanation can be extended to random forests as follows:

Definition 1. *Let $F = \{T_1, \ldots, T_m\}$ be a random forest in $\mathtt{RF}_{n,m}$, and $\boldsymbol{x} \in \{0,1\}^n$ be an instance. The* direct reason *for $\boldsymbol{x}$ given F is the term $t^F_{\boldsymbol{x}}$ defined by*

$$t^F_{\boldsymbol{x}} = \begin{cases} \bigwedge_{T_i \in F : T_i(\boldsymbol{x}) = 1} t^{T_i}_{\boldsymbol{x}} & \text{if } F(\boldsymbol{x}) = 1 \\ \bigwedge_{T_i \in F : T_i(\boldsymbol{x}) = 0} t^{T_i}_{\boldsymbol{x}} & \text{if } F(\boldsymbol{x}) = 0 \end{cases}$$

By construction, $t^F_{\boldsymbol{x}}$ is an abductive explanation for $\boldsymbol{x}$ given F that can be computed in $\mathcal{O}(n|F|)$ time.

Example 2. *Based on Example 1, consider the instance $\boldsymbol{x} = (1,1,1,1)$. Since $F(\boldsymbol{x}) = 1$, $\boldsymbol{x}$ is recognized as a Cattleya orchid. The direct reason for $\boldsymbol{x}$ given F is $t^F_{\boldsymbol{x}} = x_1 \wedge x_2 \wedge x_3 \wedge x_4$. Consider now the instance $\boldsymbol{x}' = (0,1,0,0)$, which is not recognized as a Cattleya orchid, since $F(\boldsymbol{x}) = 0$. The direct reason for $\boldsymbol{x}'$ given F is $t^F_{\boldsymbol{x}'} = x_2 \wedge \overline{x}_3 \wedge \overline{x}_4$.*

Prime-Implicant Explanations. Another valuable notion of abductive explanation is the one of *prime-implicant explanation* (Shih, Choi, and Darwiche 2018) also referred to as *sufficient reason* (Darwiche and Hirth 2020). In the setting of random forests, such explanations can be defined as follows:

Definition 2. *Let $F \in \mathtt{RF}_n$ be a random forest and $\boldsymbol{x} \in \{0,1\}^n$ be an instance. A* prime-implicant explanation *for $\boldsymbol{x}$ given F is a prime implicant t of F (resp. $\neg F$) if $F(\boldsymbol{x}) = 1$ (resp. $F(\boldsymbol{x}) = 0$) such that t covers $\boldsymbol{x}$.*

Example 3. *For our running example, $x_2 \wedge x_3 \wedge x_4$ and $x_1 \wedge x_4$ are the prime-implicant explanations for $\boldsymbol{x}$ given F. $\overline{x}_4$ and $\overline{x}_1 \wedge \overline{x}_3$ are the prime-implicant explanations for $\boldsymbol{x}'$ given F.*

Importantly, all features occurring in a prime-implicant explanation t are *relevant*. Indeed, removing any literal from t would question the fact that t implies F. Note that the direct reason $t^F_{\boldsymbol{x}}$ for $\boldsymbol{x}$ given F may contain arbitrarily many more features than a prime-implicant explanation for $\boldsymbol{x}$ given F, since this is already known in the case where F consists of a single decision tree (Izza, Ignatiev, and Marques-Silva 2020).

The problem of identifying a prime-implicant explanation t for an input instance $\boldsymbol{x} \in \{0,1\}^n$ given a random forest $F \in \mathtt{RF}_n$, has recently been shown DP-complete (Izza and Marques-Silva 2021). In fact, even the apparently simple task of *checking* whether t is an implicant of F is already hard:

Proposition 3. *Let F be a random forest in $\mathtt{RF}_n$ and t be a term over X_n. Deciding whether t is an implicant of F is* coNP*-complete.*

The above result is in stark contrast with the computational complexity of checking whether a term t is an implicant of a decision tree T. This task can be solved in polynomial time, using the fact that T can be converted (in linear time) into a clausal form $\mathtt{CNF}(T)$, together with the fact that

testing whether t implies $\texttt{CNF}(T)$ can be done in $\mathcal{O}(n|T|)$ time. That mentioned, in the case of random forests, the implicant test can be achieved via a call to a SAT oracle:

Proposition 4. *Let $F = \{T_1, \ldots, T_m\}$ be a random forest of $\texttt{RF}_{n,m}$, and t be a (satisfiable) term over X_n. Let H be the* CNF *formula*

$$\{(\overline{y}_i \vee c) : i \in [m], c \in \texttt{CNF}(\neg T_i)\} \cup \texttt{CNF}\left(\sum_{i=1}^{m} y_i > \frac{m}{2}\right)$$

where $\{y_1, \ldots, y_m\}$ are fresh variables and $\texttt{CNF}\left(\sum_{i=1}^{m} y_i > \frac{m}{2}\right)$ is a CNF *encoding of the cardinality constraint $\sum_{i=1}^{m} y_i > \frac{m}{2}$. Then, t is an implicant of F if and only if $H \wedge t$ is unsatisfiable.*

Based on such an encoding, the prime-implicant explanations for an instance $\boldsymbol{x}$ given a random forest F can be characterized in terms of MUS (minimal unsatisfiable subsets), as suggested in (Izza and Marques-Silva 2021). This characterization is useful because many SAT-based algorithms for computing a MUS (or even all MUSes) of a CNF formula have been pointed out for the past decade (Audemard, Lagniez, and Simon 2013; Liffiton et al. 2016; Marques-Silva, Janota, and Mencía 2017). They can be leveraged for computing prime-implicant explanations.

Going one step further, a natural way to improve the intelligibility of prime-implicant explanations is to focus on parsimonious ones.[2] Specifically, a *minimum-sized prime-implicant (minPI) explanation* for $\boldsymbol{x} \in \{0,1\}^n$ given $F \in \texttt{RF}_n$ is a prime-implicant explanation for $\boldsymbol{x}$ given F, that is of minimal size.

Example 4. *For our running example, $x_1 \wedge x_4$ is the unique minPI explanation for $\boldsymbol{x}$ given F, and $\overline{x}_4$ is the unique minPI explanation for $\boldsymbol{x}'$ given F.*

As a by-product of the characterization of a prime-implicant explanation in terms of MUS (Izza and Marques-Silva 2021), a minPI explanation for $\boldsymbol{x}$ given f may be viewed as a minimum-sized MUS. Thus, we can exploit algorithms for computing minimum-sized MUSes (see e.g., (Ignatiev et al. 2015)) in order to infer minPI explanations. However, identifying a minPI explanation is computationally harder than identifying a prime-implicant explanation:

Proposition 5. *Let $F \in \texttt{RF}_n$, $\boldsymbol{x} \in \{0,1\}^n$, and $k \in \mathbb{N}$. Deciding whether there exists a minPI explanation t for $\boldsymbol{x}$ given F such that t contains at most k features is Σ_2^p-complete.*

Majoritary Reasons. Based on the above considerations, a natural question arises: *does there exist a middle ground between direct reasons, which may include many irrelevant features but are easy to calculate, and prime-implicant explanations, which only contain relevant features but are potentially much harder to infer?* Inspired by the way prime implicants can be computed when dealing with decision trees, we can reply in the affirmative using the notion of *majoritary reasons*, defined as follows.

Definition 3. *Let $F = \{T_1, \ldots, T_m\}$ be a random forest in $\texttt{RF}_{n,m}$ and $\boldsymbol{x} \in \{0,1\}^n$ be an instance. A* majoritary reason *for $\boldsymbol{x}$ given F is a term t covering $\boldsymbol{x}$, such that t is an implicant of at least $\lfloor \frac{m}{2} \rfloor + 1$ decision trees T_i (resp. $\neg T_i$) if $F(\boldsymbol{x}) = 1$ (resp. $F(\boldsymbol{x}) = 0$), and for every $l \in t$, $t \setminus \{l\}$ does not satisfy this last condition.*

Example 5. *Based on our running example, the majoritary reasons for $\boldsymbol{x}$ given F are $x_1 \wedge x_2 \wedge x_4$, $x_1 \wedge x_3 \wedge x_4$, and $x_2 \wedge x_3 \wedge x_4$. Each of these explanations is more concise than the direct reason $t_{\boldsymbol{x}}^F$. For $\boldsymbol{x}'$, the majoritary reasons given F are $\overline{x}_1 \wedge \overline{x}_4$, $x_2 \wedge \overline{x}_4$, and $\overline{x}_1 \wedge x_2 \wedge \overline{x}_3$. Note that the each of the majoritary reasons $x_1 \wedge x_2 \wedge x_4$ and $x_1 \wedge x_3 \wedge x_4$ for $\boldsymbol{x}$ given F includes an irrelevant literal for the task of classifying $\boldsymbol{x}$ using F since $x_1 \wedge x_4$ is a prime-implicant explanation for $\boldsymbol{x}$ given F. Similarly, every majoritary reason for $\boldsymbol{x}'$ given F contains an irrelevant literal for the task of classifying $\boldsymbol{x}'$ using F.*

In general, the notions of majoritary reason and of prime-implicant explanation do not coincide. Indeed, a prime-implicant explanation is a prime implicant (covering $\boldsymbol{x}$) of the forest F, while a majoritary reason is an implicant t (covering $\boldsymbol{x}$) of a majority of decision trees in the forest F, satisfying the additional condition that t is a prime implicant of at least one of these decision trees.

A key observation justifying this difference is that even if every implicant of a Boolean function f is an implicant of the function $f \vee g$, it is not always the case that every *prime* implicant of f is a prime implicant of $f \vee g$. To this point, consider our running example and take the term $t = x_1 \wedge x_3 \wedge x_4$. Here, t is a majoritary reason for $\boldsymbol{x} = (1,1,1,1)$ given F, since t covers $\boldsymbol{x}$, t is a prime implicant of T_1, and t is an implicant of T_2. Thus, t is an implicant of $f = T_1 \wedge T_2$ (it is a prime one), and hence an implicant of F, using the fact that F is logically equivalent to $(T_1 \wedge T_2) \vee (T_1 \wedge T_3) \vee (T_2 \wedge T_3)$. However, t is *not* a prime implicant of F. Indeed, the sub-term $x_1 \wedge x_4$ is a prime-implicant explanation for $\boldsymbol{x}$ given F, since it is a prime implicant of F that covers $\boldsymbol{x}$.

Viewing majoritary reasons as "weak" forms of prime-implicant explanations, they can include irrelevant features:

Proposition 6. *Let $F = \{T_1, \ldots, T_m\}$ be a random forest of $\texttt{RF}_{n,m}$ and $\boldsymbol{x} \in \{0,1\}^n$. Unless $m < 3$, it can be the case that every majoritary reason for $\boldsymbol{x}$ given F contains arbitrarily many more features than any prime-implicant explanation for $\boldsymbol{x}$ given F.*

What makes majoritary reasons valuable is that they are abductive and can be generated in linear time. The evidence that any majoritary reason t for $\boldsymbol{x}$ given F is an abductive explanation comes directly from the fact that if t implies a majority of decision trees in F, then it is an implicant of F (note that the converse implication does not hold in general).

The tractability of generating majoritary reasons lies in the fact that they can be found using a simple greedy algorithm. For the case when $F(\boldsymbol{x}) = 1$, start with $t = t_{\boldsymbol{x}}$, and iterate over the literals l of t by checking whether t deprived

[2] Everything else being equal, shortest explanations can be viewed as better than longer explanations because they are easier to understand. Of course, sparsity is only one aspect of the intelligibility of an explanation. The quality of an explanation typically depends on the explainee (i.e., the person who asked for an explanation) (Doshi-Velez and Kim 2017).

of l is an implicant of at least $\lfloor \frac{m}{2} \rfloor + 1$ decision trees of F. If so, remove l from t and proceed to the next literal. Once all literals in $t_{\boldsymbol{x}}$ have been examined, the final term t is by construction an implicant of a majority of decision trees in F, such that removing any literal from it would lead to a term that is no longer an implicant of this majority. So, t is by construction a majoritary reason. The case when $F(\boldsymbol{x}) = 0$ is similar, by simply replacing each T_i with its negation in F. This greedy algorithm runs in $\mathcal{O}(n|F|)$ time, using the fact that, on each iteration, checking whether t is an implicant of T_i (for each $i \in [m]$) can be done in $\mathcal{O}(n|T_i|)$ time.

By analogy with minimum-sized prime-implicant explanations, a natural way of improving the quality of majoritary reasons is to seek for the most parsimonious ones. Formally, a *minimum-sized majoritary (minMAJ) reason* for an instance $\boldsymbol{x} \in \{0,1\}^n$ given a random forest $F \in \mathtt{RF}_n$ is a majoritary reason for $\boldsymbol{x}$ given F, that is of minimal size.

Example 6. *For our running example, the three majoritary reasons for $\boldsymbol{x}$ given F are minMAJ reasons. Contrastingly, among the majoritary reasons for $\boldsymbol{x}'$ given F, only $\overline{x}_1 \wedge \overline{x}_4$ and $x_2 \wedge \overline{x}_4$ are minMAJ reasons.*

Unsurprisingly, the optimization task for majoritary reasons is more demanding than the generation task. Still, minMAJ reasons are easier to identify than minPI reasons:

Proposition 7. *Let $F \in \mathtt{RF}_n$, $\boldsymbol{x} \in \{0,1\}^n$, and $k \in \mathbb{N}$. Deciding whether there exists a minMAJ reason t for $\boldsymbol{x}$ given F such that t contains at most k features is* NP*-complete.*

A common approach for handling NP-optimization problems is to rely on modern constraint solvers. From this perspective, recall that a PARTIAL MAXSAT problem consists of a pair $(C_{\text{soft}}, C_{\text{hard}})$ where C_{soft} and C_{hard} are (finite) sets of clauses. The goal is to find a Boolean assignment that maximizes the number of clauses c in C_{soft} that are satisfied, while satisfying all clauses in C_{hard}.

Proposition 8. *Let $F \in \mathtt{RF}_{n,m}$ and $\boldsymbol{x} \in \{0,1\}^n$. Let $(C_{\text{soft}}, C_{\text{hard}})$ be an instance of the* PARTIAL MAXSAT *problem such that:*

$$C_{\text{soft}} = \{\overline{x}_i : x_i \in t_{\boldsymbol{x}}\} \cup \{x_i : \overline{x}_i \in t_{\boldsymbol{x}}\}$$
$$C_{\text{hard}} = \{(\overline{y}_i \vee c_{|\boldsymbol{x}}) : i \in [m], c \in \mathtt{CNF}(T_i^{\pm})\} \cup \mathtt{CNF}\left(\sum_{i=1}^{m} y_i > \frac{m}{2}\right)$$

where $c_{|\boldsymbol{x}} = c \cap t_{\boldsymbol{x}}$ is the restriction of c to the literals in $t_{\boldsymbol{x}}$, $\{y_1, \ldots, y_m\}$ are fresh variables, $T_i^{\pm} = T_i$ ($i \in [m]$) if $F(\boldsymbol{x}) = 1$, $T_i^{\pm} = \neg T_i$ if $F(\boldsymbol{x}) = 0$, and $\mathtt{CNF}(\sum_{i=1}^{m} y_i > \frac{m}{2})$ is a CNF *encoding of the constraint $\sum_{i=1}^{m} y_i > \frac{m}{2}$. Let $\boldsymbol{z}^*$ be an optimal solution of $(C_{\text{soft}}, C_{\text{hard}})$. Then, $t_{\boldsymbol{x}} \cap t_{\boldsymbol{z}^*}$ is a minMAJ reason for $\boldsymbol{x}$ given F.*

Thanks to this characterization result, one can leverage the numerous algorithms that have been developed so far for PARTIAL MAXSAT (see e.g. (Ansótegui, Bonet, and Levy 2013; Morgado, Ignatiev, and Marques-Silva 2014; Narodytska and Bacchus 2014; Saikko, Berg, and Järvisalo 2016)) in order to compute minMAJ reasons.

Experiments

Experimental Setup. The empirical protocol we considered was as follows. We have focused on 15 datasets for binary classification, which are standard benchmarks from the repositories Kaggle (www.kaggle.com), OpenML (www.openml.org), or UCI (archive.ics.uci.edu/ml/). These datasets are *compas*, *placement*, *recidivism*, *adult*, *ad_data*, *mnist38*, *mnist49*, *gisette*, *dexter*, *dorothea*, *farm-ads*, *higgs_boson*, *christine*, *gina*, and *bank*. *mnist38* and *mnist49* are subsets of the *mnist* dataset, restricted to the instances of 3 and 8 (resp. 4 and 9) digits. Additional information about the datasets, (especially the number and types of features, the number of instances), and about the random forests that have been trained (especially, the number of Boolean features used, the number of trees, the depth of the trees, the mean accuracy) are available at www.cril.univ-artois.fr/expekctation/.

Categorical features have been treated as arbitrary numbers (the scale is nominal). As to numeric features, no data preprocessing has taken place: these features have been binarized on-the-fly by the random forest learning algorithm. For this learner, we have used the version 0.23.2 of the Scikit-Learn library (Pedregosa et al. 2011). The maximal depth of any decision tree in a forest has been bounded at 8. All other hyper-parameters of the learning algorithm have been set to their default value, except the number of trees. We made some preliminary tests for tuning this parameter in order to ensure that the accuracy was good enough.

For every benchmark b, a 10-fold cross validation process has been achieved: a set of 10 random forests have been computed and evaluated from the labelled instances of b, partitioned into 10 parts. One part was used as the test set and the remaining 9 parts as the training set for generating a forest. The classification performance on b was measured using the mean accuracy obtained over the 10 random forests. For each benchmark b, each random forest F, and a pool of 25 instances $\boldsymbol{x}$ drawn at random from the test set (leading to 250 instances per dataset), we have run the algorithms described in the previous section for deriving the direct reason for $\boldsymbol{x}$ given F, a prime-implicant explanation for $\boldsymbol{x}$ given F, a majoritary reason for $\boldsymbol{x}$ given F, a minPI reason for $\boldsymbol{x}$ given F, and a minMAJ reason for $\boldsymbol{x}$ given F.

For computing prime-implicant explanations and minMAJ reasons, we took advantage of the Pysat library (Ignatiev, Morgado, and Marques-Silva 2018) (version 0.1.6.dev15) that provides the implementation of the RC2 PARTIAL MAXSAT solver and an interface to MUSER (Belov and Marques-Silva 2012). For majoritary reasons, we picked up uniformly at random 50 permutations of the literals describing the instance and tried to eliminate those literals (within the greedy algorithm) following the ordering corresponding to the permutation. We kept a smallest explanation among those derived (of course, the corresponding runtime that has been measured is the cumulated time over the 50 tries). Prime-implicant explanations have been computed using MUSes, as explained before.

We also derived a "LIME explanation" for each instance. Such an explanation has been inferred as follows. Given an input instance $\boldsymbol{x}$, we first used LIME (Ribeiro, Singh, and

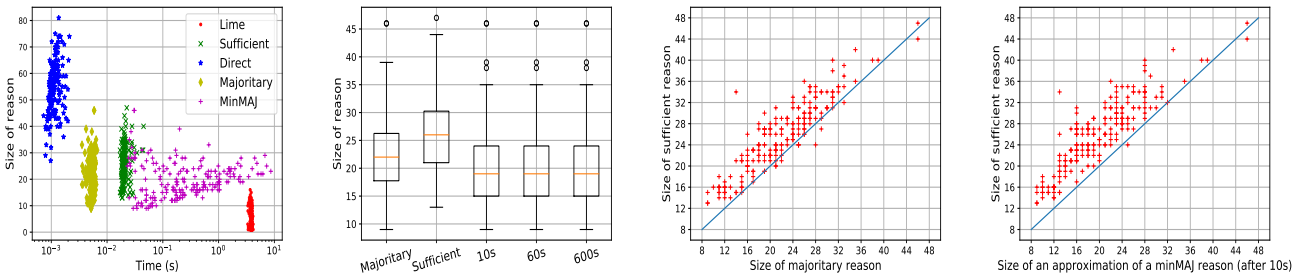

Figure 2: Empirical results for the *placement* dataset.

Guestrin 2016) to generate a linear zero-threshold function $\boldsymbol{w_x} \in \mathbb{R}^n$. The value $\boldsymbol{w_x}(\boldsymbol{z})$ of $\boldsymbol{w_x}$ on any instance $\boldsymbol{z}$ is given by $\boldsymbol{w_x}(\boldsymbol{z}) = 1$ if $\boldsymbol{w_x}^\top \boldsymbol{z} > 0$, and $\boldsymbol{w_x}(\boldsymbol{z}) = 0$ otherwise. Now, when $\boldsymbol{x}$ is classified positively by $\boldsymbol{w_x}$, in order to derive an explanation, it is enough to sum in a decreasing way the positive weights occurring in $\boldsymbol{w_x}$ until this sum exceeds (the opposite of) the sum of all the negative weights occurring in $\boldsymbol{w_x}$. The term t composed of the variables x_i associated with the positive weights which have been selected is, by construction, a minPI reason for $\boldsymbol{x}$ given $\boldsymbol{w_x}$ since for every $\boldsymbol{x'}$ covered by t, the inequality $\boldsymbol{w_x}^\top \boldsymbol{x'} > 0$ holds. Indeed, the inequality $\boldsymbol{w_x}^\top \boldsymbol{x'} > 0$ holds in the worst situation when all the variables associated with a positive weight in $\boldsymbol{w_x}$ and not belonging to t are set to 0, whilst all the variables associated with a negative weight in $\boldsymbol{w_x}$ are set to 1 (see also (Marques-Silva et al. 2020)). Instances that are classified negatively can be handled in a similar way.

Experiments have been conducted on a computer equipped with Intel(R) XEON E5-2637 CPU @ 3.5 GHz and 128 Gib of memory. A time-out (TO) of 600s has been considered for each instance and each type of explanation, except LIME ones (no time bound has been used for them).

Experimental Results. A first conclusion that can be drawn from our experiments is the intractability of computing minPI reasons in practice; this coheres with the complexity result given by Proposition 5. Thus, we have been able to compute within the time limit of 600s a minPI reason for only 10 instances and a single dataset (*compas*).

Due to space limitations, we report hereafter empirical results about three datasets only, namely *placement*, *gisette* and *dorothea*. The results obtained on the other datasets are similar and available on line. *placement* is a small dataset about the placement of 215 students in a campus; students are described using 13 features, related to their curricula, the type and work experience, and the salary. An instance is labelled positive when the student gets a job. The random forests consist of 25 trees, and their mean accuracy is 97.6%. *gisette* is much larger, including 5000 features and 7000 examples. Each feature is a pixel or a combination of pixels, and the task is to separate the digits 4 and 9. The random forests consist of 85 trees, and their mean accuracy is 96%. Finally, *dorothea* is a high-dimensional dataset, with 100,000 features and 1950 examples. Each instance is an organic molecule, and the goal is to discriminate binding compounds from non-binding ones. Here, the random forest consists of 71 trees, with a mean accuracy of 93%.

Figure 2 provides the results obtained for *placement*, using four plots. Each dot represents an instance. The first plot shows the time needed to compute a reason on the x-axis, and the size of this reason on the y-axis. On this plot, there are no dots for minPI reasons, because their computation did not terminate before the time-out. The plot also highlights that all other reasons have been computed within the time limit, and in general using a small amount of time. In particular, it shows that the direct reason can be quite large, that the computation of LIME explanations is usually more expensive than the ones of the other explanations, and that LIME explanations can be very short.[3] A box plot about the sizes of all the explanations is reported (the LIME ones and the direct reasons are not presented for the sake of readability). The figure also provides two scatter plots, aiming to compare the sizes of majoritary reasons with the sizes of prime-implicant explanations, and the sizes of the minMAJ reasons with the sizes of prime-implicant explanations.

These plots clearly show the benefits with respect to size reduction that can be offered by considering majoritary reasons and minMAJ reasons instead of prime-implicant explanations. At first sight, these empirical results may look surprising since, by construction, for any majoritary reason t for $\boldsymbol{x}$ given F (including the shortest ones) there exists at least one prime-implicant explanation for $\boldsymbol{x}$ given F that is implied by t (hence that cannot be larger). As to majoritary reasons, one must keep in mind that the result that is reported is a shortest reason out of a set of 50 majoritary reasons that are computed for each $\boldsymbol{x}$ (so to say, we leverage the tractability of computing such reasons to tackle the size issue). For minMAJ reasons, the PARTIAL MAXSAT algorithm used to compute them aims to minimize the size of the reason that is derived, while MUS algorithms for computing prime-implicant explanations do not focus on the size. Indeed, computing minimum-sized MUSes is much harder, as explained previously (see Proposition 5).

Figures 3 and 4 synthesize the results obtained for *gisette* and *dorothea*, respectively. Conclusions similar to those drawn for *placement* can be derived for *gisette* and *dorothea*,

[3] Recall that LIME explanations are not guaranteed to be abductive. See also (Narodytska et al. 2019) that reports some experiments about ANCHOR explanations.

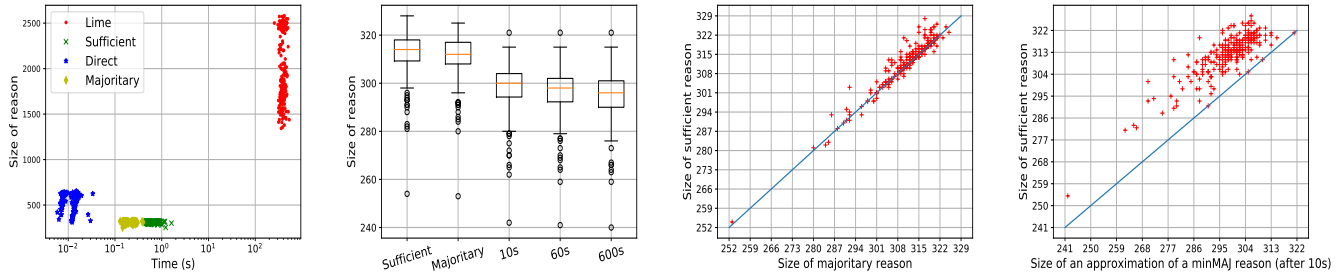

Figure 3: Empirical results for the *gisette* dataset.

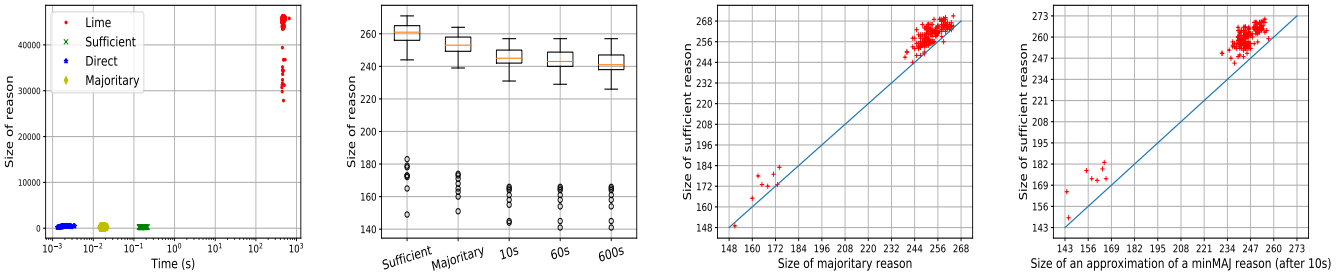

Figure 4: Empirical results for the *dorothea* dataset.

with some exceptions. First of all, no dots have been drawn for minMAJ reasons because the computation of such reasons did not terminate within the time limit. Furthermore, LIME explanations are much longer. This can be partly explained by the fact that the computation achieved by LIME relies on a binary representation of the instance that is quite different (and possibly much larger) than the one considered in the representation of the random forest. Indeed, each decision tree in the forest focuses only on a subset of most important features (in the sense of Gini criterion) found during the learning phase. In our experiments, the size of LIME explanations was typically large for high-dimensional datasets.

When minMAJ reasons are difficult to calculate (as it is the case for *gisette* and *dorothea*), a natural approach is to look for approximations. From this perspective, we took advantage of the incremental PARTIAL MAXSAT algorithm called LMHS (Saikko, Berg, and Järvisalo 2016) to do the job. Specifically, the result given in Proposition 8 provides a way to derive abductive explanations for an instance $\boldsymbol{x}$ given a random forest F in an *anytime* fashion. Basically, using LMHS, a Boolean assignment $\boldsymbol{z}$ satisfying all the hard constraints of C_{hard} and a given number, say k, of soft constraints from C_{soft} is looked for (k is set to 0 at start). If such an assignment is found, then one looks for an assignment satisfying $k+1$ soft constraint, and so on, until an optimal solution is found or a preset time bound is reached. By construction, every $\boldsymbol{z}$ that is generated that way is such that $t_{\boldsymbol{x}} \cap t_{\boldsymbol{z}}$ is an implicant of F that covers $\boldsymbol{x}$ (and hence, an abductive explanation). In practice, the approximation $\boldsymbol{z}$ of a minMAJ reason for $\boldsymbol{x}$ given F, which is obtained when the time limit is met, can be significantly shorter than the prime-implicant explanation for $\boldsymbol{x}$ given F that has been derived.

In our experiments, we used three time limits: 10s, 60s, and 600s. The results are reported in the box plots and the scatter plots in Figures 2, 3, and 4. As illustrated by the box plots, the sizes of the approximations $\boldsymbol{z}$ which are derived gently decrease with time. The scatter plots indicate that significant size savings can be achieved even for the smallest time bound of 10s that has been considered.

Conclusion

We have introduced, studied, and evaluated new notions of abductive explanations suited to random forests, namely direct reasons, majoritary reasons and minimum-sized majoritary reasons. Unlike prime-implicant explanations, majoritary reasons and their minimum-sized counterparts may contain irrelevant features. Nevertheless, in practice, majoritary reasons and minMAJ reasons appear as valuable alternative to prime-implicant explanations. Indeed, majoritary reasons can be computed in polynomial time while prime-implicant explanations cannot (unless P = NP). In addition, in most of our experiments, majoritary reasons slightly smaller than prime-implicant explanations have been computed thanks to a simple greedy algorithm with random permutations of literals. minMAJ reasons can be looked for when majoritary reasons are too large, but this is at the cost of an extra computation time that can be important, and even prohibitive in some cases. However, minMAJ reasons can be approximated using an *anytime* PARTIAL MAXSAT algorithm. Empirically, approximations can be derived within a small amount of time and their sizes are significantly smaller than the ones of prime-implicant explanations.

Acknowledgements

Many thanks to the anonymous reviewers for their comments and insights. This work has benefited from the support of the AI Chair EXPEKCTATION (ANR-19-CHIA-0005-01) of the French National Research Agency (ANR). It was also partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

References

Ansótegui, C.; Bonet, M. L.; and Levy, J. 2013. SAT-based MaxSAT algorithms. *Artificial Intelligence*, 196: 77–105.

Audemard, G.; Bellart, S.; Bounia, L.; Koriche, F.; Lagniez, J.; and Marquis, P. 2021. On the Computational Intelligibility of Boolean Classifiers. In *Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning, KR 2021, Online event, November 3-12, 2021*, 74–86.

Audemard, G.; Lagniez, J.-M.; and Simon, L. 2013. Improving Glucose for Incremental SAT Solving with Assumptions: Application to MUS Extraction. In *Proceedings of the 16th International Conference on Theory and Applications of Satisfiability Testing (SAT'13)*, 309–317.

Azar, A. T.; Elshazly, H. I.; Hassanien, A. E.; and Elkorany, A. M. 2014. A random forest classifier for lymph diseases. *Computer Methods and Programs in Biomedicine*, 113(2): 465–473.

Belov, A.; and Marques-Silva, J. 2012. MUSer2: An Efficient MUS Extractor. *J. Satisf. Boolean Model. Comput.*, 8(3/4): 123–128.

Bénard, C.; Biau, G.; Veiga, S. D.; and Scornet, E. 2021. Interpretable Random Forests via Rule Extraction. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, AISTATS'21*, 937–945.

Biau, G. 2012. Analysis of a Random Forests Model. *Journal of Machine Learning Research*, 13: 1063–1095.

Bogomolov, A.; Lepri, B.; Staiano, J.; Oliver, N.; Pianesi, F.; and Pentland, A. 2014. Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data. In *Proceedings of the 16th International Conference on Multimodal Interaction, ICMI'14*, 427–434. ACM.

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

Chen, X.; and Ishwaran, H. 2012. Random forests for genomic data analysis. *Genomics*, 99(6): 323–329.

Choi, A.; Shih, A.; Goyanka, A.; and Darwiche, A. 2020. On Symbolically Encoding the Behavior of Random Forests. In *Proceedings of the 3rd Workshop on Formal Methods for ML-Enabled Autonomous Systems (FoMLAS)*.

Criminisi, A.; and Shotton, J. 2013. *Decision Forests for Computer Vision and Medical Image Analysis*. Advances in Computer Vision and Pattern Recognition. Springer.

Cutler, R.; Thomas, C. E. J.; Beard, K. H.; Cutler, A.; Hess, K. T.; Gibson, J.; and Lawler, J. J. 2007. Random Forests for Classification in Ecology. *Ecology*, 88(11): 2783–2792.

Darwiche, A.; and Hirth, A. 2020. On the Reasons Behind Decisions. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI'20)*, 712–720.

Doshi-Velez, F.; and Kim, B. 2017. A Roadmap for a Rigorous Science of Interpretability. *CoRR*, abs/1702.08608.

Ignatiev, A.; Morgado, A.; and Marques-Silva, J. 2018. PySAT: A Python Toolkit for Prototyping with SAT Oracles. In *Proceedings of the 21st International Conference on Theory and Applications of Satisfiability Testing (SAT'2018)*, 428–437.

Ignatiev, A.; Narodytska, N.; and Marques-Silva, J. 2019. Abduction-Based Explanations for Machine Learning Models. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI'19)*, 1511–1519.

Ignatiev, A.; Previti, A.; Liffiton, M.; and Marques-Silva, J. 2015. Smallest MUS Extraction with Minimal Hitting Set Dualization. In *Proceedings of the 21st International Conference on Principles and Practice of Constraint Programming (CP'15)*, 173–182.

Izza, Y.; Ignatiev, A.; and Marques-Silva, J. 2020. On Explaining Decision Trees. *CoRR*, abs/2010.11034.

Izza, Y.; and Marques-Silva, J. 2021. On Explaining Random Forests with SAT. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI'21)*, 2584–2591.

Liffiton, M.; Previti, A.; Malik, A.; and Marques-Silva, J. 2016. Fast, flexible MUS enumeration. *Constraints An Int. J.*, 21(2): 223–250.

Marques-Silva, J.; Gerspacher, T.; Cooper, M. C.; Ignatiev, A.; and Narodytska, N. 2020. Explaining Naive Bayes and Other Linear Classifiers with Polynomial Time and Delay. In *Proc. of NeurIPS'20*.

Marques-Silva, J.; Janota, M.; and Mencía, C. 2017. Minimal sets on propositional formulae. Problems and reductions. *Artificial Intelligence*, 252: 22–50.

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.

Morgado, A.; Ignatiev, A.; and Marques-Silva, J. 2014. MSCG: Robust Core-Guided MaxSAT Solving. *J. Satisf. Boolean Model. Comput.*, 9(1): 129–134.

Narodytska, N.; and Bacchus, F. 2014. Maximum Satisfiability Using Core-Guided MaxSAT Resolution. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 2717–2723.

Narodytska, N.; Shrotri, A.; Meel, K.; Ignatiev, A.; and Marques-Silva, J. 2019. Assessing Heuristic Machine Learning Explanations with Model Counting. In *Proceedings of 22nd International Conference on the Theory and Applications of Satisfiability Testing (SAT'19)*, 267–278.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.;

Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. ACM.

Ribeiro, M.-T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-Precision Model-Agnostic Explanations. In McIlraith, S. A.; and Weinberger, K. Q., eds., *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 1527–1535.

Saikko, P.; Berg, J.; and Järvisalo, M. 2016. LMHS: A SAT-IP Hybrid MaxSAT Solver. In *Proceedings of the 19th International Conference of Theory and Applications of Satisfiability Testing (SAT'16)*, 539–546.

Shih, A.; Choi, A.; and Darwiche, A. 2018. A Symbolic Approach to Explaining Bayesian Network Classifiers. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI'18)*, 5103–5111.