On Contrastive Explanations for Tree-Based Classifiers

Gilles Audemard, Jean-Marie Lagniez, Pierre Marquis, Nicolas Szczepanski

On Contrastive Explanations for Tree-Based Classifiers

Gilles Audemard[a], Jean-Marie Lagniez[a], Pierre Marquis[a,b] and Nicolas Szczepanski[c]

[a]Univ. Artois, CNRS, CRIL, France
[b]Institut Universitaire de France, France
[c]IRT SystemX, France
ORCiD ID: Gilles Audemard https://orcid.org/0000-0003-2604-9657,
Jean-Marie Lagniez https://orcid.org/0000-0002-6557-4115,
Pierre Marquis https://orcid.org/0000-0002-7979-6608,
Nicolas Szczepanski https://orcid.org/0000-0001-7553-5657

Abstract. We define contrastive explanations that are suited to tree-based classifiers. In our framework, contrastive explanations are based on the set of (possibly non-independent) Boolean characteristics used by the classifier and are at least as general as contrastive explanations based on the set of characteristics of the instances considered at start. We investigate the computational complexity of computing contrastive explanations for Boolean classifiers (including tree-based ones), when the Boolean conditions used are not independent. Finally, we present and evaluate empirically an algorithm for computing minimum-size contrastive explanations for random forests.

1 Introduction

Explaining the behaviour of AI systems is an issue of major significance in the perspective of trustworthy AI. Thus, recent years have seen a remarkable boom in work aimed at verifying AI systems and explaining the outputs they generate (see for instance [18, 19, 21, 24, 28, 31, 34, 41, 1, 9, 39]).

Several types of explanations can be defined when dealing with AI systems that implement classifiers f, i.e., mappings from a set $\boldsymbol{X}$ of instances to a set $\mathcal{L}$ of classes (see e.g., [23]). On the one hand, *abductive explanations* are about explaining the classification of an instance $\boldsymbol{x} \in \boldsymbol{X}$ achieved by f, by focusing on a subset of the *characteristics* (i.e., the pairs attribute-value) of $\boldsymbol{x}$, that are sufficient to justify the classification $f(\boldsymbol{x})$ made, in the sense that any instance sharing this subset of characteristics is necessarily classified in the same way as $\boldsymbol{x}$. On the other hand, *contrastive explanations* for an input instance $\boldsymbol{x}$ aim to explain why $\boldsymbol{x}$ has *not* been classified by f as expected by the user who asked for an explanation, aka the explainee.

In this paper, we focus on *contrastive explanations* and *tree-based classifiers* f in the binary case (i.e., when $\mathcal{L} = \{0, 1\}$). Beyond decision trees [8, 38], our study includes random forests [7] and boosted trees [14, 40, 15]. Such models are typically more accurate than decision trees (boosted trees are among state-of-the-art ML models when dealing with tabular data [6]) but they are also more opaque [1, 2] and the combinatorial and non-differentiable nature of tree ensembles makes also the generation of explanations more challenging.

In previous work (see [17] for a survey), a contrastive explanation for $\boldsymbol{x}$ is defined as a *contrastive instance* (i.e., an instance classified in a different way than $\boldsymbol{x}$), that is as *close* as possible to $\boldsymbol{x}$. Closeness can be measured in various ways, using distances, similarities, sets of characteristics, or even sets of attributes. Thus, in [22], a contrastive explanation for $\boldsymbol{x}$ is defined as a minimal subset c of the set of attributes $\mathcal{A}$ used to describe $\boldsymbol{x}$ such that there exist a value v_i for each $A_i \in c$ and a contrastive instance $\boldsymbol{x}_c$ that takes value v_i for $A_i \in c$, $\boldsymbol{x}_c$ coincides with $\boldsymbol{x}$ on every attribute outside c, and $\boldsymbol{x}_c$ is not classified in the same way as $\boldsymbol{x}$. When all the attributes of $\mathcal{A}$ are Boolean ones, such a contrastive explanation for $\boldsymbol{x}$ can also be defined as a minimal subset of the characteristics of the input instance $\boldsymbol{x}$ that must be flipped in $\boldsymbol{x}$ in order to get an instance $\boldsymbol{x}_c$ classified in a different way. Indeed, the characteristics of $\boldsymbol{x}_c$ can be easily deduced from the set of Boolean attributes to be modified provided that $\boldsymbol{x}$ is known. Such contrastive explanations are referred to as *necessary reasons* [12].

Beyond closeness to the input instance $\boldsymbol{x}$, contrastive explanations can also be assessed by considering their *generality*, i.e., the population of feasible contrastive instances they cover. Indeed, on the one hand, the number of contrastive explanations for a given $\boldsymbol{x}$ can be huge, so that computing all of them can be out of reach. Furthermore, providing a large number of contrastive explanations to the explainee is useless most of the time since she / he will not have the cognitive capacity to grasp them as a whole. On the other hand, contrastive explanations reduced to single instances may turn out to be outliers, and not true contrastive explanations. Thus, providing instead a reduced set of more general explanations is better. However, when taking generality into account, the definitions of contrastive explanations based on $\mathcal{A}$, as considered in previous work, *are not suited to tree-based models*. Let us illustrate this issue using a very simple scenario that will serve as a running example in the paper.

Example 1. *Suppose that the decision tree classifier f, depicted on Figure 1, is used to determine whether a loan must be granted or not to an applicant. A_1 is a numerical attribute that gives the annual incomes of the applicant. A_2 is a Boolean attribute that indicates whether the applicant has already reimbursed a previous loan. Alice wants to get a loan. Alice's annual incomes are equal to $\$18k$ and Alice has not reimbursed yet a previous loan. Alice corresponds to an instance $\boldsymbol{x}^A = (18, 0)$ and $f(\boldsymbol{x}^A) = 0$. The loan is not granted, and Alice would like to know what she could do to change the decision.*

Two contrastive instances from $\boldsymbol{X}$ that are as close as possible to

* Corresponding Author. Email: marquis@cril.univ-artois.fr

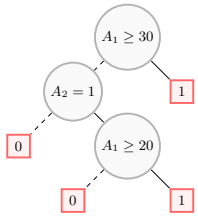

Figure 1: A simple decision tree classifier f.

$\boldsymbol{x}^A$ while being classified as positive by f are $(20, 1)$ and $(30, 0)$. Using words, "increase your incomes to $\$20k$ and reimburse your previous loan" and "increase your incomes to $\$30k$ ". Better contrastive explanations would be "increase your incomes to at least *$\$20k$ and reimburse your previous loan" and "increase your incomes to* at least *$\$30k$ ". Indeed, suppose that Alice does not know whether A_1 is monotonic for f, i.e., if $\boldsymbol{x}$ and $\boldsymbol{x}'$ are two instances that coincide except possibly on A_1, $f(\boldsymbol{x}) = 1$ and $x_1 \leq x'_1$, then $f(\boldsymbol{x}') = 1$. In this case, Alice cannot infer from the explanation $(30, 0)$ what would happen if her annual incomes increased to $\$35k$: would she get the loan as well, or not? While the predictor f gives a positive answer to this question, this is not reflected in the explanations $(20, 1)$ and $(30, 0)$ that are generated. More general explanations covering the contrastive instances $(20, 1)$ and $(30, 0)$ but also other contrastive instances would be welcome.*

Alternatively, if one considers that contrastive explanations are given as minimal subsets c of attributes to be modified in the input instance as suggested in [22], $\{A_1\}$ is the unique contrastive explanation for $\boldsymbol{x}^A$. Using words, "to get the loan, it is enough to change the value of your annual incomes". Using this definition, one gets a contrastive explanation that is not informative enough: Alice surely expects to know to which extent her annual incomes must be updated in order to get the loan. Especially, it is not the case that changing the value 18 of A_1 to any other value will lead to a contrastive instance: if Alice's annual incomes decrease (or increase but remain below $\$30k$), the loan will not be granted. Thus, $A_1 \neq 18$ covers instances that are not contrastive instances for $\boldsymbol{x}^A$. Furthermore, the contrastive instance $(20, 1)$ is not covered by it because of the subset-minimality requirement (the value of A_2 in $\boldsymbol{x}^A$ must also be updated if one wants to cover it).

Our goal in this paper is *to show how to define, characterize, and compute contrastive explanations suited to tree-based classifiers*, while avoiding the shortcomings of previous proposals, illustrated in the example above. This goes through the definition of a new set of instances $\boldsymbol{X}_f$ based on Boolean attributes given by the conditions used in f, so that every instance $\boldsymbol{x} \in \boldsymbol{X}$ can be rewritten into an instance, denoted $r_f(\boldsymbol{x})$, that belongs to $\boldsymbol{X}_f$ and is such that $f(r_f(\boldsymbol{x})) = f(\boldsymbol{x})$. On the running example, three Boolean attributes A_1^1, A_1^2, A_2^1 defined by $A_1^1 = (A_1 \geq 20)$, $A_1^2 = (A_1 \geq 30)$, and $A_2^1 = (A_2 = 1)$ used in f can be considered for describing instances and we have $r_f(\boldsymbol{x}^A) = (0, 0, 0)$. Using conjunctively-interpreted sets of characteristics instead of vectors, the instance $\boldsymbol{x}^A$ corresponding to Alice is described primarily by $\{A_1 = 18, A_2 = 0\}$ and, once rewritten, by $\{\overline{A_1^1}, \overline{A_1^2}, \overline{A_2^1}\}$ (or equivalently by $\{\overline{(A_1 \geq 20)}, \overline{(A_1 \geq 30)}, \overline{(A_2 = 1)}\}$). Accordingly, f will be viewed both as a mapping from $\boldsymbol{X}$ to $\mathcal{L}$ and, alternatively, as a mapping from $\boldsymbol{X}_f = \{r_f(\boldsymbol{x}) : \boldsymbol{x} \in \boldsymbol{X}\}$ to $\mathcal{L}$. In the latter case, f is a Boolean function since every attribute used is considered as a Boolean one. However, f may contain Boolean attributes that are not pairwise independent because they come from the same (non-Boolean) attribute A_i used to describe instances from $\boldsymbol{X}$. Thus, some propositional constraints Σ forming a domain theory indicating how the Boolean conditions used in f are logically connected must be taken into account to refrain from deriving incorrect explanations, based on instances that are not feasible. For instance, no rewritten instance from $\boldsymbol{X}_f$ can be such that A_1^2 is true and A_1^1 is false. The pair (f, Σ) is referred to as a *constrained decision-function* [16].

Our contributions are as follows. After some formal preliminaries (Section 2), we show in Section 3 that contrastive explanations for rewritten instances $r_f(\boldsymbol{x})$ must be privileged to contrastive explanations for initial instances $\boldsymbol{x}$. There are two reasons for it. On the one hand, explanations for rewritten instances are *as intelligible* as explanations based on the set of characteristics of the input instances, since their meaning is primarily based on the same attributes, those from $\mathcal{A}$. On the other hand, explanations for rewritten instances are often *more general*, so more informative and more robust than explanations represented in the initial space of characteristics. Then, we focus on contrastive explanations for rewritten instances. In Section 4, we define (weak, subset-minimal, and minimum-size) contrastive explanations for instances based on the set of characteristics of f given a constrained decision-function (f, Σ). We show how those explanations can be characterized in terms of (prime) implicates. We identify the computational complexity of recognizing such explanations and contrast it with the complexity of recognizing abductive explanations given a constrained decision-function (such abductive explanations have been considered in [16]). Recognizing contrastive explanations appears as "mildly" hard (first level of the polynomial hierarchy), which suggests that their computation is feasible in practice in many cases. To evaluate it, we describe in Section 5 an approach to derive minimum-size contrastive explanations and present some empirical results showing that this approach can be used in practice.

A folder containing a full-proof version of the paper, a more detailed description of the datasets, and the code used in our experiments is available online at http://www.cril.univ-artois.fr/expekctation/; this code is also part of our XAI library `PyXAI` (https://www.cril.univ-artois.fr/pyxai/).

2 Preliminaries

Classification Let $\mathcal{A} = \{A_1, \ldots, A_k\}$ be a finite set of attributes (aka features), where each attribute is Boolean, categorical (aka nominal), or numerical. The *domain* D_i of A_i ($i \in [k]$) is $\{0, 1\}$ when A_i is Boolean, a finite set of values that are not ordered when A_i is categorical (for instance $D_i = \{orange, white, green\}$), and (typically) $D_i = \mathbb{N}$ or $\mathbb{R}$ when A_i is numerical. We note $\mathcal{A}_{boo}$ (resp. $\mathcal{A}_{cat}$, $\mathcal{A}_{num}$) the subset of $\mathcal{A}$ consisting of Boolean (resp. categorical, numerical) attributes.

An *instance* $\boldsymbol{x}$ over $\mathcal{A}$ is a tuple from $D_1 \times \ldots \times D_k$. Every $\boldsymbol{x} = (v_1, \ldots, v_k)$ is also viewed logically as the conjunctively-interpreted set $t_{\boldsymbol{x}}$ of Boolean conditions (alias characteristics) $\{(A_i = v_i) : i \in [k]\}$. $\boldsymbol{X}$ is the set of all instances. A *binary classifier* f over $\mathcal{A}$ is a mapping from $\boldsymbol{X}$ to $\mathcal{L} = \{0, 1\}$. An instance $\boldsymbol{x} \in \boldsymbol{X}$ is *positive* when $f(\boldsymbol{x}) = 1$ and it is *negative* when $f(\boldsymbol{x}) = 0$.

A *decision tree* over $\mathcal{A}$ is a binary tree T, each of whose internal nodes is a decision node, labeled with a Boolean condition on $A_i \in \mathcal{A}$, and each leaf is labeled by an element of $\mathcal{L}$. Whenever A_i is numerical, the set of Boolean conditions labelling the nodes over A_i used in f takes the form $(A_i \geq v_j^i)$. Whenever A_i is categorical and it has been one-hot encoded, the set of Boolean conditions labelling

the nodes over A_i used in f takes the form $(A_i = v_j^i)$. In both cases, the set of encountered values v_j^i in those nodes forms a subset D_i^f of the domain D_i of A_i, and D_i^f is not a singleton in general. 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 Boolean condition labelling the node is evaluated to 0 (resp. 1) for $\boldsymbol{x}$.

A *random forest* over $\mathcal{A}$ is an ensemble $F = \{T_1, \cdots, T_m\}$, where each T_i ($i \in [m]$) is a decision tree over $\mathcal{A}$, 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.

Boolean functions 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. A Boolean vector $\boldsymbol{x} \in \{0,1\}^n$ is a *model* of f if $f(\boldsymbol{x}) = 1$. Otherwise, $\boldsymbol{x}$ is a *counter-model* of f. $[f]$ denotes the set of all models of f.

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. f is *valid* when it has no negative instance. If f and g are two propositional formulae over X_n, f *entails* g, noted $f \models g$, if and only if $[f] \subseteq [g]$ holds and f and g are *equivalent*, noted $f \equiv g$, if and only if $[f] = [g]$. A *literal* over a variable $x \in X_n$ is x itself (a positive literal) or its negation $\neg x$, also denoted $\overline{x}_i$ (a negative literal). L_{X_n} is the set of all literals over X_n. A *term* t is a conjunction of literals, and a *clause* c is a disjunction of literals. In what follows, we often treat instances as terms, and terms as sets of literals. For an assignment $\boldsymbol{z} \in \{0,1\}^n$, the corresponding *canonical term* is $t_{\boldsymbol{z}} = \bigwedge_{i=1}^{n} x_i^{z_i}$ where $x_i^0 = \overline{x}_i$ and $x_i^1 = x_i$. A term t *covers* an assignment $\boldsymbol{x}$ if $t \subseteq t_{\boldsymbol{x}}$. A satisfiable term t is an *implicant* of f if and only if $t \models f$ holds, and t is a *prime implicant* of f if and only if t is an implicant of f and no proper subset of t is an implicant of f. A non-valid clause c is an *implicate* of f if and only if $f \models c$ holds, and c is a *prime implicate* of f if and only if c is an implicate of f and no proper subset of c is an implicate of f. 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 $Var(f)$.

When every Boolean condition occurring in a decision tree T (resp. a random forest F) is viewed as a Boolean variable, T (resp. F) can be viewed as a Boolean function over X_n. The class of decision trees over X_n is denoted $\mathtt{DT}_n$, and the class of random forests over X_n is denoted $\mathtt{RF}_n$.

3 Improving Generality by Rewriting Instances

Two families of contrastive explanations for instances from $\boldsymbol{X}$ can be considered when f is a tree-based classifier, given that two distinct spaces of characteristics can be used for describing instances:

Definition 1. *Given a finite set of attributes $\mathcal{A}$ and a tree-based binary classifier f over A:*

- *The* space of characteristics of the instances *is the set* $\mathcal{C} = \{(A_i = v_j^i) : A_i \in \mathcal{A}, v_j^i \in D_i\}$.
- *The* space of characteristics of the classifier *is the set of literals over the Boolean conditions $\mathcal{C}_f$ of the form $(A_i = v_j^i)$ and $(A_i \geq v_j^i)$ that are used in f.*

Accordingly, instances $\boldsymbol{x}$ to be explained are either considered as they were *given primarily* (conjunctions of characteristics, i.e., of pairs attribute-value, from $\mathcal{C}$), or they are first *rewritten into conjunctions of literals $r_f(\boldsymbol{x})$ over the Boolean variables $\mathcal{C}_f$ used by the predictor f*. The rewrite function r_f we consider is thus a mapping from $\boldsymbol{X}$ to the terms over $\mathcal{C}_f$.

We argue that rewritten instances should be considered as first-class candidates because contrastive explanations for instances represented using $\mathcal{C}$ are unnecessarily specific or not informative enough (as shown before, using the running example). In order to prove that rewritten instances are more general than instances considered at start, we first show how $r_f(\boldsymbol{x})$ is logically connected to $\boldsymbol{x}$ via a *domain theory*. Formally, for every numerical attribute A_i, one assumes an implicit First-Order Logic (FOL) theory capturing the semantics of $=$ and $\leq$ over the set of numbers in the domain of A_i is implicitly taken into account (e.g., the theory used is DLO – Dense Linear Order – if the values of the attribute are real numbers). In the categorical case, one makes the *unique name assumption*: if v_p, v_q are two distinct values in the domain of A_i, then $(A_i = v_p)$ implies that $\overline{(A_i = v_q)}$.

Instead of the FOL theory itself, when dealing with an instance $\boldsymbol{x} \in \boldsymbol{X}$, it is enough to consider the *propositional grounding* of the theory given f and $\boldsymbol{x} = (v_1, \ldots, v_k)$. Using the propositional grounding of the theory instead of the theory itself is more convenient from a computational perspective. For any numerical attribute $A_i \in \mathcal{A}_{num}$, let $D_f(A_i) = \{v_1^i, \ldots, v_{p_i}^i\}$ be the set of values – ordered in ascending way (i.e., $v_1^i < \ldots, < v_{p_i}^i$) – about A_i that can be found in the decision nodes of f. The corresponding propositional grounding is the formula

$$\Sigma_{num}(A_i, f, \boldsymbol{x}) = ((A_i = v_i) \Rightarrow c(v_i, D_f(A_i))) \wedge \Sigma_{num}(A_i, f)$$

where $\Sigma_{num}(A_i, f) = \bigwedge_{j=1}^{p_i - 1}(\overline{(A_i \geq v_j^i)} \Rightarrow \overline{(A_i \geq v_{j+1}^i)})$ and $c(v_i, D_f(A_i)) = (\overline{(A_i \geq v_i \uparrow)} \wedge (A_i \geq v_i \downarrow)$ with $v_i \uparrow = min(\{v_j^i \in D_f(A_i) : v_i < v_j^i\})$ and $v_i \downarrow = max(\{v_j^i \in D_f(A_i) : v_j^i \leq v_i\})$, when $v_1^i \leq v^i < v_{p_i}^i$, $c(v_i, D_f(A_i)) = (A_i \geq v_{p_i}^i)$ when $v^i \geq v_{p_i}^i$, and $c(v_i, D_f(A_i)) = \overline{(A_i \geq v_1^i)}$ when $v^i < v_1^i$.

Similarly, for any categorical attribute $A_i \in \mathcal{A}_{cat}$, let $D_f(A_i) = \{v_1^i, \ldots, v_{p_i}^i\}$ be the set of values that can be found in the decision nodes of f. The corresponding propositional grounding is given by the formula

$$\Sigma_{cat}(A_i, f, \boldsymbol{x}) = ((A_i = v_i) \Rightarrow \bigwedge_{v_j^i \in D_f(A_i) \setminus \{v_i\}} \overline{(A_i = v_j^i)}) \wedge \Sigma_{cat}(A_i, f)$$

where $\Sigma_{cat}(A_i, f) = \bigwedge_{j=1}^{p_i - 1} \bigwedge_{l=j+1}^{p_i}((A_i = v_j^i) \Rightarrow \overline{(A_i = v_l^i)})$. In $\Sigma_{num}(A_i, f, \boldsymbol{x})$ and $\Sigma_{cat}(A_i, f, \boldsymbol{x})$, $(A_i = v_i)$ and $(A_i = v_j^i)$ ($j \in [p_i]$) are viewed as propositional variables.

Whatever the type of A_i (numerical or categorical), the corresponding grounding is composed of two parts (that are connected conjunctively): a first part that depends on $\boldsymbol{x}$ (and more precisely of the value v_i taken by A_i in $\boldsymbol{x}$) and a second part only about the Boolean conditions used by f. This second part is denoted by $\Sigma_{num}(A_i, f)$ when A_i is numerical and by $\Sigma_{cat}(A_i, f)$ when A_i is categorical. We denote by $\Sigma(f)$ the conjunction $\bigwedge_{A_i \in \mathcal{A}_{num}} \Sigma_{num}(A_i, f) \wedge \bigwedge_{A_i \in \mathcal{A}_{cat}} \Sigma_{cat}(A_i, f)$.

We are now in position to make more formal the notion of rewritten instance.

Definition 2. *Let $\boldsymbol{x} = (v_1, \ldots, v_k) \in \boldsymbol{X}$ be an instance over $\mathcal{A} = \{A_1, \ldots, A_k\}$ where $(A_i = v_i) \in \mathcal{C}$ ($i \in [k]$). Let f be a tree-based classifier over $\mathcal{A}$. The* rewritten instance *$r_f(\boldsymbol{x})$ over $\mathcal{C}_f$ is given by $t_{r_f(\boldsymbol{x})}$ where $t_{r_f(\boldsymbol{x})}$ is the set of all literals over $\mathcal{C}_f$ that are logical consequences of $t_{\boldsymbol{x}}$ given*

$$\Sigma(f, \boldsymbol{x}) = \bigwedge_{A_i \in \mathcal{A}_{num}} \Sigma_{num}(A_i, f, \boldsymbol{x}) \wedge \bigwedge_{A_i \in \mathcal{A}_{cat}} \Sigma_{cat}(A_i, f, \boldsymbol{x}).$$

By definition, $t_{r_f(\boldsymbol{x})}$ is a logical consequence of $t_{\boldsymbol{x}}$ given $\Sigma(f, \boldsymbol{x})$. It is also easy to check that $r_f(\boldsymbol{x})$ satisfies the underlying theory $\Sigma(f, \boldsymbol{x})$. Especially, for every $\boldsymbol{x} \in \boldsymbol{X}$, $r_f(\boldsymbol{x})$ satisfies $\Sigma(f)$.

In the general case, $t_{r_f(\boldsymbol{x})}$ is *not equivalent* given $\Sigma(f, \boldsymbol{x})$ to $t_{\boldsymbol{x}}$ but is *strictly more general*. Thus, on the running example,

$$t_{r_f(\boldsymbol{x}^A)} = \{\overline{(A_1 \geq 30)}, \overline{(A_1 \geq 20)}, \overline{(A_2 = 1)}\}$$

captures not only Alice but *the whole population of instances $\boldsymbol{x} \in \boldsymbol{X}$* with less than $\$20k$ annual incomes and a previous loan not reimbursed. Since contrastive instances are instances, the gain of generality obtained by considering rewritten instances also applies to contrastive explanations. Note by the way that "more general" does not imply "shorter" contrastive explanations in the general case (this holds only when explanations are based on the same set of characteristics). Indeed, rewritten instances are usually longer than the instances considered at start, so this also applies to contrastive instances. This can be easily explained by the fact that several thresholds v_j^i can be considered in the tree-based classifier f for the same numerical attribute A_i from $\mathcal{C}$. Of course, it can be the case that an attribute A_i from $\mathcal{C}$ is detected as useless by f (on the example, a numerical attribute A_3 indicating the level of qualifications of the applicant could be considered in $\mathcal{C}$ but not be used in the predictor f since it appears as irrelevant to discriminate the positive instances from the negative ones). In such a case, A_i does not correspond to any Boolean attribute in $\mathcal{C}_f$.

Because Boolean conditions in $\mathcal{C}_f$ are connected when they are issued from the same attribute $A_i \in \mathcal{A}$, $\Sigma(f)$ must be taken into account to discard explanations that do not comply with $\Sigma(f)$, and are not legit as a consequence [44]. As a matter of illustration, consider Alice's case again. $c = \{(A_1 \geq 30)\}$ is a subset-minimal contrastive explanation for $\boldsymbol{x}$ given f in the sense of [22]. However, no instance from $\boldsymbol{X}$ matches this representation in $\mathcal{C}_f$ because $t_{r_f(\boldsymbol{x}^A)_c} = \{(A_1 \geq 30), \overline{(A_1 \geq 20)}, \overline{(A_2 = 1)}\}$ conflicts with $\Sigma(f) = \overline{(A_1 \geq 20)} \Rightarrow \overline{(A_1 \geq 30)}$. In order to eliminate such impossible instances, $\Sigma(f)$ must be taken into account in the definition of contrastive explanations. The next section indicates how to do it.

4 On Contrastive Explanations

In the following, we define notions of contrastive explanations suited to classifiers based on *Boolean variables that are logically connected* by a domain theory Σ, and we investigate their computational complexity. The proposed setting covers the case of tree-based classifiers f involving numerical or categorical attributes and rewritten instances as discussed in the previous section (in this case, we take $\Sigma = \Sigma(f)$), but is actually more general. For instance, hierarchical attributes (i.e., categorical attributes connected into an ontology) could be considered as well in this setting. Classifiers based on Boolean variables that are logically connected are referred to as *constrained decision-functions* in [16].

Definition 3. *[16] Let $X_n = \{x_1, \ldots, x_n\}$ be a set of Boolean variables. A* constrained decision-function *over X_n is a pair (f, Σ) where $f \in \mathcal{F}_n$ and Σ is a propositional formula over X_n. Σ indicates how the Boolean variables from X_n are logically connected.*

Contrastive explanations given a constrained decision-function Given a constrained decision-function, the next definition introduces notions of contrastive explanation, subset-minimal contrastive explanation, and minimum-size contrastive explanation for an instance.

Definition 4. *Let (f, Σ) be a constrained decision-function and $\boldsymbol{x} \in [\Sigma]$ be an instance.*

- *A* contrastive explanation *for $\boldsymbol{x}$ given (f, Σ) is a set $c \subseteq t_{\boldsymbol{x}}$ such that the vector $\boldsymbol{x}_c \in \{0, 1\}^n$ that coincides with $\boldsymbol{x}$ except on the characteristics of c is such that $\boldsymbol{x}_c \in [\Sigma]$ and $f(\boldsymbol{x}_c) \neq f(\boldsymbol{x})$.*
- *A* subset-minimal contrastive explanation *for $\boldsymbol{x}$ given (f, Σ) is a contrastive explanation c for $\boldsymbol{x}$ given (f, Σ) such that no proper subset of c is a contrastive explanation for $\boldsymbol{x}$ given (f, Σ).*
- *A* minimum-size contrastive explanation *for $\boldsymbol{x}$ given (f, Σ) is a contrastive explanation c for $\boldsymbol{x}$ given (f, Σ) such that no contrastive explanation c' for $\boldsymbol{x}$ given (f, Σ) such that $|c'| < |c|$ exists.*

Those notions of contrastive explanations echo the following notions of abductive explanations:

Definition 5. *Let (f, Σ) be a constrained decision-function and $\boldsymbol{x} \in [\Sigma]$ be an instance s.t. $f(\boldsymbol{x}) = 1$ (resp. $f(\boldsymbol{x}) = 0$).*

- *An* abductive explanation *for $\boldsymbol{x}$ given (f, Σ) is a set $t \subseteq t_{\boldsymbol{x}}$ such that $t \wedge \Sigma \models f$ (resp. $t \wedge \Sigma \models \overline{f}$).*
- *A* subset-minimal abductive explanation *for $\boldsymbol{x}$ given (f, Σ) is an abductive explanation t for $\boldsymbol{x}$ given (f, Σ) such that no proper subset of t is an abductive explanation for $\boldsymbol{x}$ given (f, Σ).*
- *A* minimum-size abductive explanation *for $\boldsymbol{x}$ given (f, Σ) is an abductive explanation t for $\boldsymbol{x}$ given (f, Σ) such that no abductive explanation t' for $\boldsymbol{x}$ given (f, Σ) such that $|t'| < |t|$ exists.*

Subset-minimal abductive explanations (alias sufficient reasons) given a constrained decision-function have been investigated in [16] (see also [10] for the case when the domain theory encodes numerical attributes). Subset-minimal abductive explanations are connected to subset-minimal contrastive explanations via a minimal hitting set duality [22] that still holds when a domain theory Σ is taken into account [44], and that also corresponds, in logical terms, to the well-known duality between prime implicants and prime implicates.

Our characterization results take advantage of this duality and related results from [16] and [12], as well as the notion of *universal literal quantification* considered in [13]. Let us recall this notion. If ℓ is a literal over x, then the *universal quantification* of ℓ from f, noted $\forall \ell \cdot f$, is the formula $(\ell \vee (f|\overline{\ell})) \wedge (f|\ell)$. In this expression, $(f|\ell)$ denotes the *conditioning* of f by ℓ. If $\ell = x$ is a positive literal (resp. $\ell = \overline{x}$ is a negative literal) , $(f|\ell)$ is the formula obtained by replacing in f every occurrence of x by 1 (resp. 0). When t is a set of literals, $\forall t \cup \{\ell_{k+1}\} \cdot f$ denotes the formula $\forall t \cdot (\forall \ell_{k+1} \cdot f)$. Finally, $\forall \boldsymbol{x}$ is a short for $\forall t_{\boldsymbol{x}}$.

We are now ready to present the following characterization results for contrastive explanations. Note that while the running example is focused on domain theories used to properly encode numerical and categorical attributes, all the propositions reported in this section apply to *any constrained decision-function*, and thus may concern *more general domain theories*.

Proposition 1. *Let (f, Σ) be a constrained decision-function and $\boldsymbol{x} \in [\Sigma]$ be an instance s.t. $f(\boldsymbol{x}) = 1$ (resp. $f(\boldsymbol{x}) = 0$).*

- *The contrastive explanations for $\boldsymbol{x}$ given (f, Σ) are the sets of literals c such that $\bigvee_{\ell \in c} \ell$ is an implicate of $\forall \boldsymbol{x} \cdot (\Sigma \Rightarrow f)$ (resp. $\forall \boldsymbol{x} \cdot (\Sigma \Rightarrow \overline{f})$).*
- *The subset-minimal contrastive explanations for $\boldsymbol{x}$ given (f, Σ) are the sets of literals c such that $\bigvee_{\ell \in c} \ell$ is a prime implicate of $\forall \boldsymbol{x} \cdot (\Sigma \Rightarrow f)$ (resp. $\forall \boldsymbol{x} \cdot (\Sigma \Rightarrow \overline{f})$).*
- *The minimum-size contrastive explanations for $\boldsymbol{x}$ given (f, Σ) are the sets of literals c such that $\bigvee_{\ell \in c} \ell$ is a minimum-size prime implicate of $\forall \boldsymbol{x} \cdot (\Sigma \Rightarrow f)$ (resp. $\forall \boldsymbol{x} \cdot (\Sigma \Rightarrow \overline{f})$).*

Let us illustrate this proposition using Alice's example. We have

$$t_{r_f(\boldsymbol{x}^A)} = \{\overline{(A_1 \geq 30)}, \overline{(A_1 \geq 20)}, \overline{(A_2 = 1)}\}.$$

f is equivalent to $(A_1 \geq 30) \vee ((A_1 \geq 20) \wedge (A_2 = 1))$ and

$$\Sigma = \Sigma(f) = \overline{(A_1 \geq 30)} \vee (A_1 \geq 20).$$

Thus, $\Sigma \Rightarrow \overline{f}$ is equivalent to $\overline{(A_1 \geq 20)} \vee (\overline{(A_1 \geq 30)} \wedge \overline{(A_2 = 1)})$ and $\forall r_f(\boldsymbol{x}^A) \cdot (\Sigma \Rightarrow \overline{f})$ is equivalent to

$$\overline{(A_1 \geq 20)} \vee (\overline{(A_1 \geq 30)} \wedge \overline{(A_2 = 1)}).$$

This formula has two prime implicates:

$$\overline{(A_1 \geq 20)} \vee \overline{(A_1 \geq 30)} \text{ and } \overline{(A_1 \geq 20)} \vee \overline{(A_2 = 1)}.$$

Accordingly, $r_f(\boldsymbol{x}^A)$ has two subset-minimal contrastive explanations given (f, Σ), namely

$$c_1 = \{\overline{(A_1 \geq 20)}, \overline{(A_1 \geq 30)}\} \text{ and } c_2 = \{\overline{(A_1 \geq 20)}, \overline{(A_2 = 1)}\}.$$

They are also minimum-size contrastive explanations. They correspond respectively to the contrastive instances given by

$$t_{r_f(\boldsymbol{x}^A)_{c_1}} = \{(A_1 \geq 30), (A_1 \geq 20), \overline{(A_2 = 1)}\}, \text{ and}$$

$$t_{r_f(\boldsymbol{x}^A)_{c_2}} = \{\overline{(A_1 \geq 30)}, (A_1 \geq 20), (A_2 = 1)\}.$$

While it provides a simple, logic-based, characterization of contrastive explanations given a constrained decision-function, Proposition 1 does not ensure that the computation of the set of all contrastive explanations for a instance given a constrained decision-function is feasible. This is not the case in general, due to the intrinsic difficulty of deriving (subset-minimal or minimum-size) contrastive explanations (that will be discussed next) but also to the number of explanations. Indeed, in the unconstrained case (i.e., when Σ is valid – e.g., $\Sigma = 1$), an instance $\boldsymbol{x}$ can have exponentially many (minimum-size, thus subset-minimal) contrastive explanations given a random forest.

Proposition 2. *Let $F = \{T_1, \cdots, T_m\}$ be a random forest of $\mathtt{RF}_n$ and $\boldsymbol{x} \in \{0,1\}^n$ be an instance. The number of minimum-size contrastive explanations for $\boldsymbol{x}$ given $(F, 1)$ can be exponential in the number n of attributes and in the number m of trees used in F.*

However, Proposition 1 can be exploited to reason about the whole set of (subset-minimal or minimum-size) contrastive explanations for $\boldsymbol{x}$ without needing to enumerate the elements of the set. For instance, we can take advantage of it to derive the necessary (resp. relevant) characteristics of subset-minimal contrastive explanations, i.e., those characteristics occurring in all (resp. at least one) subset-minimal contrastive explanation(s) [3]. In particular, when $f(r_f(\boldsymbol{x})) = 0$, those characteristics are given by the literals implying $\forall r_f(\boldsymbol{x}) \cdot (\Sigma \Rightarrow \overline{f})$ (resp. the literals $\forall r_f(\boldsymbol{x}) \cdot (\Sigma \Rightarrow \overline{f})$ depends on [29]). Thus, on Alice's example, $\overline{(A_1 \geq 20)}$ is the unique necessary characteristic of subset-minimal contrastive explanations for $r_f(\boldsymbol{x}^A)$, while all the characteristics in $t_{r_f(\boldsymbol{x}^A)}$ are relevant.

The complexity of contrastive explanations Despite the duality linking them, contrastive explanations differ from abductive explanations on several aspects when it comes to their computation. First of all, while an instance $\boldsymbol{x} \in [\Sigma]$ always has an abductive explanation given (f, Σ) (indeed, $t_{\boldsymbol{x}}$ is such an abductive explanation), $\boldsymbol{x}$ does not always have a contrastive explanation given (f, Σ). To be more precise:

Proposition 3. *Let (f, Σ) be a constrained decision-function and $\boldsymbol{x} \in [\Sigma]$ be an instance such that $f(\boldsymbol{x}) = 1$ (resp. $f(\boldsymbol{x}) = 0$). $\boldsymbol{x}$ has a contrastive explanation given (f, Σ) if and only if $\neg f \wedge \Sigma$ (resp. $f \wedge \Sigma$) is satisfiable. Deciding whether $\boldsymbol{x}$ has a contrastive explanation given (f, Σ) is* NP*-complete.* NP*-hardness still holds when f is represented by a random forest from $\mathtt{RF}_n$ and $\Sigma = 1$.*

Another significant difference is based on the fact that recognizing contrastive explanations is computationally easier than recognizing abductive explanations, i.e., subsets of $t_{\boldsymbol{x}}$ that are implicants of $\Sigma \Rightarrow f$ (this last problem is coNP-complete in general, and even in the restricted case when f is represented by a random forest from $\mathtt{RF}_n$ and $\Sigma = 1$ [4]).

Proposition 4. *Let (f, Σ) be a constrained decision-function and $\boldsymbol{x} \in [\Sigma]$ be an instance. Let $c \subseteq t_{\boldsymbol{x}}$. Deciding whether c is a contrastive explanation for $\boldsymbol{x}$ given (f, Σ) is in* P*.*

Contrastingly, recognizing (subset-minimal, or even minimum-size) contrastive explanations is intractable:

Proposition 5. *Let (f, Σ) be a constrained decision-function and $\boldsymbol{x} \in [\Sigma]$ be an instance. Let $c \subseteq t_{\boldsymbol{x}}$. Deciding whether c is a subset-minimal contrastive explanation for $\boldsymbol{x}$ given (f, Σ) is* coNP*-complete.* coNP*-hardness still holds when f is represented by a random forest from $\mathtt{RF}_n$ and $\Sigma = 1$.*

Proposition 6. *Let (f, Σ) be a constrained decision-function and $\boldsymbol{x} \in [\Sigma]$ be an instance. Let $c \subseteq t_{\boldsymbol{x}}$. Deciding whether c is a minimum-size contrastive explanation for $\boldsymbol{x}$ given (f, Σ) is* coNP*-complete.* coNP*-hardness still holds when f is represented by a random forest from $\mathtt{RF}_n$ and $\Sigma = 1$.*

Subset-minimal and minimum-size abductive explanations are harder to recognize, even in the unconstrained case (i.e., when Σ is valid). Indeed, in the case of a random forest $F \in \mathtt{RF}_n$, deciding whether $t \subseteq t_{\boldsymbol{x}}$ is a subset-minimal abductive explanation for $\boldsymbol{x}$ given F has been shown DP-complete [25] (the membership to DP extends to the general case when the classifier is any Boolean function $f \in \mathcal{F}_n$). In [4], it has been shown that, given $\boldsymbol{x} \in \{0,1\}^n$, $F \in \mathtt{RF}_n$ such that $F(\boldsymbol{x}) = 1$, and an integer k, deciding whether there exists a minimum-size abductive explanation t for $\boldsymbol{x}$ given F such that $|t| \leq k$ is Σ_2^p-complete. On this basis, one can show that deciding whether t is a minimum-size abductive explanation for $\boldsymbol{x}$ given F is Π_2^p-complete. Thus, under the assumption that the polynomial hierarchy does not collapse, identifying a minimum-size abductive explanation for an instance given a random forest is computationally harder than identifying a minimum-size contrastive explanation for an instance given a random forest. Notably, under the same assumption, identifying a minimum-size abductive explanation for an instance given a decision tree is also computationally harder than identifying a minimum-size contrastive explanation for an instance given a decision tree (indeed, the former is NP-hard [5], while the latter can be done in (deterministic) polynomial time (see e.g., [20]).

5 Computing Minimum-Size Contrastive Explanations

Leveraging Proposition 4, one can easily design an algorithm for computing a *minimum-size* (thus subset-minimal) contrastive explanation for $\boldsymbol{x}$ given (f, Σ) using a PARTIAL MAXSAT solver. By definition, such contrastive explanations are less numerous than the contrastive explanations for $\boldsymbol{x}$ given (f, Σ) and they express smaller changes (in terms of the number of characteristics of $\boldsymbol{x}$ to be modified).

Proposition 7. *Let (f, Σ) be a constrained decision-function and $\boldsymbol{x} \in [\Sigma]$ be an instance such that $f(\boldsymbol{x}) = 1$.*[1] *Let $(C_{\rm soft}, C_{\rm hard})$ be an instance of the* PARTIAL MAXSAT *problem such that $C_{\rm soft} = t_{\boldsymbol{x}}$ and $C_{\rm hard} = CNF(\Sigma \wedge \overline{f})$ where $CNF(\Sigma \wedge \overline{f})$ is a* `CNF` *encoding of $\Sigma \wedge \overline{f}$. Let $\boldsymbol{z}^*$ be an optimal solution of $(C_{\rm soft}, C_{\rm hard})$. Then, $c = t_{\boldsymbol{x}} \setminus t_{\boldsymbol{z}^*}$ is a minimum-size contrastive explanation for $\boldsymbol{x}$ given (f, Σ) and we have $t_{\boldsymbol{x}_c} = t_{\boldsymbol{z}^*} \cap L_{X_n}$.*

Let us recall that an instance of PARTIAL MAXSAT consists of a pair $(C_{\rm soft}, C_{\rm hard})$ where $C_{\rm soft}$ and $C_{\rm hard}$ are (finite) sets of clauses. According to Proposition 7, finding a minimum-size contrastive explanation c for $\boldsymbol{x}$ given f then mainly amounts to finding an assignment $\boldsymbol{z}$ of the propositional variables involved in $C_{\rm soft} \cup C_{\rm hard}$ that maximizes the number of clauses in $C_{\rm soft}$ that are satisfied, while satisfying all clauses in $C_{\rm hard}$. Here $CNF(\Sigma \wedge \overline{f})$ denotes any `CNF` formula that is query-equivalent to $\Sigma \wedge \overline{f}$, i.e., equivalent to it over X_n. Such a `CNF` formula can be generated in linear time from the representation of f using Tseitin transformation [43] or Plaisted/Greenbaum one [37] – those transformations require new variables to be introduced. Those new variables are dummy ones. Notably, they are not involved in the minimum-size contrastive explanations that are generated.

Clearly enough, one can take advantage of the PARTIAL MAXSAT characterization above for generating a preset number of minimum-size contrastive explanations via the use of *blocking clauses*. Basically, the approach is as follows: one generates a first minimum-size contrastive explanation c, then one adds to $C_{\rm hard}$ the negation of the corresponding contrastive instance $t_{\boldsymbol{x}_c}$ as a clause and we resume until the bound is reached or no solution exists or the size of the last explanation that has been generated is strictly larger than the size of the first explanation c that has been computed.

Empirical evaluation We have run some experiments in order to assess to which extent minimum-size contrastive explanations can be computed in practice.

Setting Our empirical protocol was as follows. We have focused on 20 datasets (some of them based on numerical, categorical, and Boolean attributes) 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/). In these datasets, the number of attributes (features) varies from 4 to 20000, and the number of instances from 170 to 32561.

For each dataset, random forest classifiers F have been learned using the Scikit-Learn library (version 1.0.2) [36]. Categorical features have been one-hot encoded. Numerical features, have been binarized on-the-fly by the random forest learning algorithm. The domain theory Σ used was $\Sigma(F)$. All hyper-parameters of the learning algorithm have been set to their default value (in particular, 100 trees per forest and no depth bound on the trees). Indeed, the approach to XAI we follow, though exact (we do not approximate the predictor F nor the explanations), is *post-hoc*, i.e., it is intended to take place once F has been learned, whatever its accuracy. As it can be observed, using default parameters led to quite a good accuracy for almost all datasets under consideration (see Table 1), but not in every case (especially, for the `breast Tumor` dataset), and again, this is on purpose since our objective is to derive explanations suited to predictors *as they are*, and not as they could have been. What we want to evaluate is the ability to derive explanations in practice even if the accuracy of the predictor is rather low.

For every dataset, a 10-fold cross validation process has been achieved. For each dataset, each random forest F, and a pool of 10 instances $\boldsymbol{x}$ over $\mathcal{C}_f$ drawn at random from the test set and satisfying Σ (leading to 100 instances per dataset), we have run our algorithm for computing a minimum-size contrastive explanation c for $\boldsymbol{x}$ given (F, Σ). For making the experiments, we took advantage of the PyXAI library (https://www.cril.univ-artois.fr/pyxai/) and used the `openwbo` (WEIGHTED) PARTIAL MAXSAT solver [32].

For each dataset, we counted the number of instances (out of 100) for which a minimum-size contrastive explanation has been computed in due time (a time-out (TO) of 100s has been considered per instance). For each instance for which the computation has been successful, we measured the time needed to get the result and the size of the resulting contrastive explanation. All the experiments have been conducted on a computer equipped with Intel(R) XEON E5-2637 CPU @ 3.5 GHz and 128 Gib of memory.

Results A synthesis of the results obtained is provided in Table 1. The columns give, from left to right, the name of the dataset, the number of attributes $\mathcal{A}$ its is based on, the mean number of Boolean variables from $\mathcal{C}_F$ used in the random forests that have been generated, the mean number of attributes from $\mathcal{A}$ used in the random forests that have been generated, the mean accuracy of the random forests. The two remaining groups of columns are about the performance of our algorithm for computing minimum-size contrastive explanations c, in terms of run time and size of the explanations, respectively. The first group reports successively the mean run times in seconds and the number of TOs met by our algorithm. The last group focuses on the size of c, and gives successively the number of literals over $\mathcal{C}_F$ in c, the number of attributes from $\mathcal{A}$ the literals of c are issued from, the percentage of literals in c relative to $|\mathcal{C}_F|$, and finally the percentage of number of attributes from $\mathcal{A}$ the literals of c are issued, relative to $|\mathcal{A}_{used}|$. When measuring the sizes of the explanations, we exploited the fact that `openwbo` exhibits an *anytime* behaviour: when the algorithm timed out, a contrastive explanation (that is not of minimal size in general) can be derived nevertheless. Hence its size can be considered in the statistics drawn.

The results show that our algorithm to derive minimum-size contrastive explanations has been successful most of the time. TOs were met frequently only for the largest datasets (`mnist38`, `christine`, `gisette` and `dexter`), involving hundreds attributes and leading to random forests based on more than 7500 Boolean conditions. When no TOs occurred (value 0 in column **#TO** of Table 1), each of the 100 contrastive explanations that have been derived for the dataset under consideration is guaranteed to be of minimal size. Seemingly, the accuracy of the predictor does not have any impact on the time needed by our algorithm for deriving a minimum-size explanation. Especially, for the datasets for which the computation always terminated in due time, minimum-size contrastive explanations have been derived in a short amount of time (in average, 13.55s).

We also ran additional experiments with a larger time-out (1200s per instance). For each instance for which a contrastive explanation

[1] If $\boldsymbol{x}$ is such that $f(\boldsymbol{x}) = 0$, then consider $\overline{f}$ instead of f.

Dataset	$\lvert\mathcal{A}\rvert$	F			Run time		Size			
		$\lvert\mathcal{C}_F\rvert$	$\lvert\mathcal{A}_{used}\rvert$	Accuracy	Time	#TO	$\lvert\mathcal{C}_F\rvert$	$\lvert\mathcal{A}\rvert$	$\%\lvert\mathcal{C}_F\rvert$	$\%\lvert\mathcal{A}\rvert$
balance-scale	4	28.0(±0.0)	4.0(±0.0)	91.2(±3.6)	4.23(±0.9)	0	5.28(±2.8)	1.51(±0.5)	18.86(±9.8)	37.75(±12.5)
bupa	5	254.9(±18.7)	5.0(±0.0)	97.39(±2.4)	0.07(±0.0)	0	54.91(±25.3)	2.55(±0.6)	21.52(±9.6)	51.0(±12.4)
compas	7	69.5(±1.0)	7.0(±0.0)	66.57(±1.9)	73.55(±15.1)	0	4.6(±6.3)	1.0(±0.0)	6.64(±9.1)	14.29(±0.0)
breastTumor	9	117.8(±1.6)	9.0(±0.0)	53.88(±10.4)	2.24(±0.4)	0	11.66(±17.6)	1.24(±0.5)	9.88(±14.9)	13.78(±5.5)
contraceptive	9	112.8(±1.0)	9.0(±0.0)	66.6(±3.4)	20.42(±0.7)	0	13.26(±14.4)	1.04(±0.2)	11.76(±12.8)	11.55(±2.2)
cleveland	13	556.5(±11.9)	13.0(±0.0)	80.91(±6.4)	0.44(±0.1)	0	54.29(±64.5)	1.88(±0.9)	9.77(±11.7)	14.46(±6.6)
adult	13	50718.2(±205.5)	13.0(±0.0)	85.44(±0.8)	29.86(±2.7)	0	1343.29(±4043.2)	1.06(±0.2)	2.65(±8.0)	8.15(±1.8)
australian	14	1437.2(±30.2)	14.0(±0.0)	86.81(±4.9)	2.0(±0.8)	0	91.72(±116.5)	1.58(±0.6)	6.35(±8.0)	11.29(±4.4)
bank	16	5200.1(±48.1)	16.0(±0.0)	89.69(±1.9)	9.97(±0.2)	0	1175.5(±696.9)	1.43(±0.6)	22.61(±13.4)	8.94(±3.9)
melb	17	28380.8(±65.9)	17.0(±0.0)	92.09(±0.6)	10.94(±0.6)	0	1845.7(±1081.3)	1.56(±0.6)	6.5(±3.8)	9.17(±3.3)
german	19	521.4(±23.5)	19.0(±0.0)	96.3(±2.6)	20.32(±23.3)	3	$\leq$120.02(±86.1)	$\leq$5.64(±2.1)	23.13(±16.7)	$\leq$29.69(±11.0)
default-paiement	23	173268.3(±524.0)	23.0(±0.0)	81.61(±0.7)	27.43(±18.8)	27	$\leq$2859.03(±10196.2)	$\leq$2.83(±3.6)	1.65(±5.9)	$\leq$12.31(±15.7)
biodegradation	41	5730.7(±89.3)	40.9(±0.3)	87.78(±2.9)	4.57(±9.0)	1	$\leq$459.52(±344.8)	$\leq$2.66(±2.8)	8.02(±6.0)	$\leq$6.51(±6.8)
divorce	54	116.5(±8.8)	50.4(±1.6)	97.65(±3.9)	3.34(±7.9)	0	15.59(±10.1)	5.57(±3.3)	13.39(±8.6)	11.04(±6.6)
spambase	57	15005.5(±91.7)	57.0(±0.0)	95.41(±0.6)	5.91(±8.6)	0	517.69(±451.9)	1.81(±0.9)	3.45(±3.0)	3.17(±1.6)
mnist38	784	32638.6(±173.4)	545.8(±3.2)	98.7(±0.4)	37.94(±31.0)	85	$\leq$1882.97(±895.6)	$\leq$54.38(±26.9)	5.77(±2.7)	$\leq$9.96(±4.9)
cnae	856	555.1(±13.2)	520.0(±12.1)	99.54(±0.6)	1.47(±0.4)	0	2.14(±1.1)	1.85(±0.7)	0.39(±0.2)	0.36(±0.1)
christine	1636	43587.5(±113.3)	1605.1(±1.0)	72.13(±2.2)	13.0(±17.1)	69	$\leq$221.45(±160.0)	$\leq$19.54(±14.4)	0.51(±0.4)	$\leq$1.22(±0.9)
gisette	5000	24464.6(±172.0)	4107.2(±16.1)	97.53(±0.6)	0.87(±0.0)	97	$\leq$274.02(±105.7)	$\leq$48.41(±17.1)	1.12(±0.4)	$\leq$1.18(±0.4)
dexter	20000	7892.9(±60.1)	3452.4(±32.5)	93.83(±2.4)	15.33(±24.2)	47	$\leq$63.91(±48.0)	$\leq$13.57(±13.7)	0.81(±0.6)	$\leq$0.39(±0.4)

Table 1: Performance of our algorithm for computing minimum-size contrastive explanations in terms of run time and size.

of minimal size has been derived within 1200s, we have computed the difference between the size of the explanation obtained after 100s and the optimal size. This difference in average was quite small (the largest value was 11.2 for gisette), showing that the quality of the explanations obtained after 100s is pretty good in average.

Our experiments also show that the sizes of the minimum-size contrastive explanations c that are derived can be large enough (even we consider only the number of attributes from $\mathcal{A}$ the literals of c are issued from), and possibly too large to be understood as a whole by a human user (the limit is usually set to 7 ± 2 [33]). Does it mean that the computation of such explanations is useless in this case? For sure, no! Once again, our perspective is not to invent short explanations when they do not exist but to explain the behaviour of the classifier as it is, and not as it could be. If the explanations that are generated do not sufficiently comply with the user's expectations, he/she is free not to trust in the corresponding prediction. Finally, it can be observed that minimum-size contrastive explanations are in practice quite small *relative to* the instances $\boldsymbol{x}$ one started with. Changing the values of a few percentage (in average, 12.81%) of the attributes of $\mathcal{A}$ used in F is enough to change the way $\boldsymbol{x}$ is classified.

6 Other Related Work

As sketched in the introduction, many works about the generation of contrastive explanations focus on the computation of a nearest contrastive instance (see [26, 35, 17] for recent references) using optimization techniques (e.g., MILP). Various distances / norms over $\boldsymbol{X}$ have been taken into account and constraints have been considered as well to discard instances that cannot be used as contrastive explanations because they are impossible, cannot be reached because their derivation would involve non-actionable attributes, or are viewed as outliers. Approximation of the splits of the decision trees is sometimes used in order to recover a differentiable setting [30]. Approximately nearest contrastive instances with a preset degree of accuracy can also be considered [27]. Heuristic approaches based on small alterations of the paths in the decision trees in order to change the decision made have been proposed as well [42].

In our approach, instances are considered over $\mathcal{C}_f$, thus described using Boolean attributes. In such a case, contrastive explanations correspond to *sets of contrastive instances* over $\mathcal{C}$, hence they are typically *more general* than single contrastive instances. The minimum-size contrastive explanations for a rewritten instance $r_f(\boldsymbol{x})$ are the (provably) nearest contrastive instances of $r_f(\boldsymbol{x})$ over $\mathcal{C}_f$ w.r.t. Manhattan/Hamming distance (or, equivalently, ℓ_1-norm). No commensurability assumptions about the scales of numerical attributes of $\mathcal{A}$ is needed in our approach, while the difficult task of identifying meaningful scaling factors has to be achieved when local distances over D_is must be aggregated to define a distance over $\boldsymbol{X}$ (e.g., how far is $(50, 30)$ to $(25, 20)$ when x_1 is the age of the applicant and x_2 his/her income in $\$k$?).

Closer to our work is [11], where binary variables denoting the fact that a numerical attribute from $\mathcal{A}$ takes its values within a specific interval are used. Contrastive explanations based on those variables are generated. As in our approach, such explanations correspond (in general) not to a single instance of $\boldsymbol{X}$ but to a population of instances. The cost function used in this work is parameterized by a cost matrix that can take into account the characteristics of the instance $\boldsymbol{x}$ and different norms, including the ℓ_0-norm over $\mathcal{C}$, which corresponds to the ℓ_1-norm over $\mathcal{C}_f$ as in our work. Domain theories for numerical attributes are considered implicitly in the encoding pointed out in [11], while our approach takes advantage of any explicit domain theory in `CNF` format Besides, the encoding used in [11] requires to introduce numerous binary variables, especially one variable per leaf of each tree, while our approach is far less demanding in this respect (the binary variables used correspond basically to the conditions found in the trees). Finally, as the other works mentioned above, [11] does not consider the issue of identifying the complexity of recognizing a contrastive explanation for constrained decision-functions, which makes it significantly different from our own work.

7 Conclusion

When dealing with tree-based classifiers f, instances $\boldsymbol{x}$ can be considered either as they are or alternatively, as instances $r_f(\boldsymbol{x})$ rewritten using the Boolean conditions appearing in f. In this paper, we have shown that contrastive explanations for rewritten instances are valuable since they are more general. We have defined notions of contrastive explanations given a constrained decision-function (f, Σ) and pointed out characterizations in terms of (prime) implicates. We have identified the computational complexity of recognizing contrastive explanations. An approach to derive minimum-size contrastive explanations has also been presented, and experiments have shown that this approach can be used in practice for deriving explanations for random forests based on hundreds Boolean conditions.

Acknowledgements

The authors would like to thank 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

[1] A. Barredo Arrieta, N. Díaz R., J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, 'Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI', *Inf. Fusion*, **58**, 82–115, (2020).

[2] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J.-M. Lagniez, and P. Marquis, 'On the computational intelligibility of boolean classifiers', in *Proc. of KR'21*, pp. 74–86, (2021).

[3] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J.-M. Lagniez, and P. Marquis, 'On the explanatory power of boolean decision trees', *Data Knowl. Eng.*, **142**, 102088, (2022).

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

[5] P. Barceló, M. Monet, J. Pérez, and B. Subercaseaux, 'Model interpretability through the lens of computational complexity', in *Proc. of NeurIPS'20*, (2020).

[6] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, 'Deep neural networks and tabular data: A survey', *CoRR*, **abs/2110.01889**, (2021).

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

[8] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, 1984.

[9] R. Caruana, S. M. Lundberg, M. Túlio Ribeiro, H. Nori, and S. Jenkins, 'Intelligible and explainable machine learning: Best practices and practical challenges', in *Proc. of KDD'20*, pp. 3511–3512. ACM, (2020).

[10] A. Choi, A. Shih, A. Goyanka, and A. Darwiche, 'On symbolically encoding the behavior of random forests', in *Proc. of the 3rd Workshop on Formal Methods for ML-Enabled Autonomous Systems (FoMLAS)*, (2020).

[11] Z. Cui, W. Chen, Y. He, and Y. Chen, 'Optimal action extraction for random forests and boosted trees', in *Proc. of KDD'15*, pp. 179–188, (2015).

[12] A. Darwiche and C. Ji, 'On the computation of necessary and sufficient explanations', in *Proc. of AAAI'22*, pp. 5582–5591, (2022).

[13] A. Darwiche and P. Marquis, 'On quantifying literals in Boolean logic and its applications to explainable AI', *J. Artif. Intell. Res.*, **72**, 285–328, (2021).

[14] Y. Freund and R.E. Schapire, 'A decision-theoretic generalization of on-line learning and an application to boosting', *J. Comput. Syst. Sci.*, **55**(1), 119–139, (1997).

[15] J. H. Friedman, 'Greedy function approximation: A gradient boosted machine', *The Annals of Statistics*, **29**(5), 1189–1232, (2001).

[16] N. Gorji and S. Rubin, 'Sufficient reasons for classifier decisions in the presence of domain constraints', in *Proc. of AAAI'22*, pp. 5660–5667, (2022).

[17] R. Guidotti, 'Counterfactual explanations and how to find them: literature review and benchmarking', *Data Mining and Knowledge Discovery*, (2022).

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

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

[20] X. Huang, Y. Izza, A. Ignatiev, and J. Marques-Silva, 'On efficiently explaining graph-based classifiers', in *Proc. of KR'21*, pp. 356–367, (2021).

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

[22] A. Ignatiev, N. Narodytska, N. Asher, and J. Marques-Silva, 'From contrastive to abductive explanations and back again', in *Proc. of AIxIA 2020 - Advances in Artificial Intelligence, Revised Selected Papers*, volume 12414 of *LNCS*, pp. 335–355, (2020).

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

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

[25] Y. Izza and J. Marques-Silva, 'On explaining random forests with SAT', in *Proc. of IJCAI'21)*, pp. 2584–2591, (2021).

[26] K. Kanamori, T. Takagi, K. Kobayashi, and H. Arimura, 'DACE: distribution-aware counterfactual explanation by mixed-integer linear optimization', in *Proc. of IJCAI'20*, pp. 2855–2862, (2020).

[27] AH. Karimi, G. Barthe, B. Balle, and I. Valera, 'Model-agnostic counterfactual explanations for consequential decisions', in *Proc. of AISTATS'20*, pp. 895–905, (2020).

[28] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, 'Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)', in *Proc. of ICML'18*, pp. 2668–2677, (2018).

[29] J. Lang, P. Liberatore, and P. Marquis, 'Propositional independence - formula-variable independence and forgetting', *Journal of Artificial Intelligence Research*, **18**, 391–443, (2003).

[30] A. Lucic, H. Oosterhuis, H. Haned, and M. de Rijke, 'FOCUS: flexible optimizable counterfactual explanations for tree ensembles', in *Proc. of AAAI'22*, pp. 5313–5322, (2022).

[31] S. Lundberg and S-I. Lee, 'A unified approach to interpreting model predictions', in *Proc. of NIPS'17*, pp. 4765–4774, (2017).

[32] R. Martins, V. M. Manquinho, and I. Lynce, 'Open-wbo: A modular maxsat solver,', in *Proc. of SAT*, pp. 438–445, (2014).

[33] G. A. Miller, 'The magical number seven, plus or minus two: Some limits on our capacity for processing information', *The Psychological Review*, **63**(2), 81–97, (1956).

[34] Ch. Molnar, *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable*, Leanpub, 2019.

[35] A. Parmentier and T. Vidal, 'Optimal counterfactual explanations in tree ensembles', in *Proc. of ICML'21*, volume 139 of *Proceedings of Machine Learning Research*, (2021).

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research*, **12**, 2825–2830, (2011).

[37] D. A. Plaisted and S. Greenbaum, 'A structure-preserving clause form translation', *Journal of Symbolic Computation*, **2**(3), 293–304, (1986).

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

[39] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, 'Interpretable machine learning: Fundamental principles and 10 grand challenges', *CoRR*, **abs/2103.11251**, (2021).

[40] R.E. Schapire and Y. Freund, *Boosting: Foundations and Algorithms*, MIT Press, 2014.

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

[42] G. Tolomei, F. Silvestri, A. Haines, and M. Lalmas, 'Interpretable predictions of tree-based ensembles via actionable feature tweaking', in *Proc. of KDD'17*, pp. 465–474, (2017).

[43] G.S. Tseitin, *On the complexity of derivation in propositional calculus*, chapter Structures in Constructive Mathematics and Mathematical Logic, 115–125, Steklov Mathematical Institute, 1968.

[44] J. Yu, A. Ignatiev, P. J. Stuckey, N. Narodytska, and J. Marques-Silva, 'Eliminating the impossible, whatever remains must be true', *CoRR*, **abs/2206.09551**, (2022).