On Contrastive Explanations for Tree-Based Classifiers  
(extended version with more details and proofs)

Gilles Audemard\textsuperscript{a}, Jean-Marie Lagniez\textsuperscript{a}, Pierre Marquis\textsuperscript{a,b} and Nicolas Szczepanski\textsuperscript{c}

\textsuperscript{a}Univ. Artois, CNRS, CRIL, France  
\textsuperscript{b}Institut Universitaire de France, France  
\textsuperscript{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 [20, 21, 23, 26, 30, 34, 37, 44, 1, 11, 42]). In previous work (see [19] for a survey), a contrastive explanation for \(x\) is defined as a contrastive instance (i.e., an instance classified in a different way than \(x\)), that is as close as possible to \(x\). Closeness can be measured in various ways, using distances, similarities, sets of characteristics, or even sets of attributes. Thus, in [24], a contrastive explanation for \(x\) is defined as a minimal subset \(c\) of the set of attributes \(A\) used to describe \(x\) such that there exist a value \(v_i\) for each \(A_i \in c\) and a contrastive instance \(x_c\) that takes value \(v_i\) for \(A_i \in c\), coincides with \(x\) on every attribute outside \(c\), and \(x_c\) is not classified in the same way as \(x\). When all the attributes of \(A\) are Boolean ones, such a contrastive explanation for \(x\) can also be defined as a minimal subset of the characteristics of the input instance \(x\) that must be flipped in \(x\) in order to get an instance \(x_c\) classified in a different way. Indeed, the characteristics of \(x_c\) can be easily deduced from the set of Boolean attributes to be modified provided that \(x\) is known. Such contrastive explanations are referred to as necessary reasons [14].

Beyond closeness to the input instance \(x\), 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 \(x\) 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 \(A\), 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 $81k and...
Alice has not reimbursed yet a previous loan. Alice corresponds to an instance \( x^A = (18, 0) \) and \( f(x^A) = 0 \). The loan is not granted, and Alice would like to know what she could do to change the decision.

Two contrastive instances from \( X \) that are as close as possible to \( x^A \) while being classified as positive by \( f \) are \((20, 1)\) and \((30, 0)\). Using words, “increase your incomes to at least $20k and reimburse your previous loan” and “increase your incomes to at least $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 \( x \) and \( x' \) are two instances that coincide except possibly on \( A_1 \), \( f(x) = 1 \) and \( x_1 \leq x'_1 \), then \( f(x') = 1 \). 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 [24], \( \{A_1\} \) is the unique contrastive explanation for \( x^A \). 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 of 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 \( x^A \). Furthermore, the contrastive instance \((20, 1)\) is not covered by it because of the subset-minimality requirement (the value of \( A_2 \) in \( x^A \) 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 \( X_f \) based on Boolean attributes given by the conditions used in \( f \), so that every instance \( x \in X \) can be rewritten into an instance, denoted \( r_f(x) \), that belongs to \( X_f \) and is such that \( f(r_f(x)) = f(x) \). On the running example, three Boolean attributes \( A_1, A_2, A_3 \) defined by \( A_1 = (\geq 20), A_2 = (\geq 30), A_3 = (\geq 1) \) used in \( f \) can be considered for describing instances and we have \( r_f(x^A) = (0, 0, 0) \). Using conjunctively-interpreted sets of characteristics instead of vectors, the instance \( x^A \) corresponding to Alice is described primarily by \( \{A_1 = 18, A_2 = 0\} \) and, once rewritten, by \( \{A_1, A_3\} \) (equivalently by \( \{(A_1 \geq 20), (A_2 \geq 30), (A_3 = 1)\} \)). Accordingly, \( f \) will be viewed both as a mapping from \( X \) to \( \mathbb{L} \) and, alternatively, as a mapping from \( X_f = \{r_f(x) : x \in X\} \) to \( \mathbb{L} \). 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 \( X \). Thus, some propositional constraints \( \Sigma \) 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 \( X_f \) can be such that \( A_1 = 30 \) and \( A_2 = 20 \) is false. The pair \( (f, \Sigma) \) is referred to as a constrained decision-function [18].

Our contributions are as follows. After some formal preliminaries (Section 2), we show in Section 3 that contrastive explanations for rewritten instances \( r_f(x) \) must be privileged to contrastive explanations for initial instances \( x \). 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 \( A \). 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 \( (f, \Sigma) \). 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 [18]). 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.

Proofs are reported in a final appendix. A folder containing a more detailed description of the datasets, and the code used in our experiments is available online at http://www.cril.univ-artois.fr/expectcation/; this code is also part of our XAI library PyXAI (https://www.cril.univ-artois.fr/pyxai/).

2 Preliminaries

Classification Let \( A = \{A_1, \ldots, A_k\} \) 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 \( D_i = \{\text{orange, white, green}\} \)), and (typically) \( D_i = \mathbb{N} \) or \( \mathbb{R} \) when \( A_i \) is numerical. We note \( A_{boa} \) (resp. \( A_{cat}, A_{num} \)) the subset of \( A \) consisting of Boolean (resp. categorical, numerical) attributes.

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

A decision tree over \( A \) is a binary tree \( T \), each of whose internal nodes is a decision node, labeled with a Boolean condition on \( A_i \in \{
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 $(A_i \geq v_i)$. 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 $(A_i = v_i)$. In both cases, the set of encountered values $v_i$ in those nodes forms a subset $D_i$ of the domain $D_i$ of $A_i$, and $D_i$ is not a singleton in general. The value $T(x)$ of $T$ on an input instance $x$ 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 $x$.

A random forest over $A$ is an ensemble $F = \{T_1, \ldots, T_m\}$, where each $T_i (i \in [m])$ is a decision tree over $A$, and such that the value $F(x)$ is given by

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

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 $F_X$, we denote the class of all Boolean functions from $\{0,1\}^n$ to $\{0,1\}$, and we use $X_n = \{x_1, \ldots, x_n\}$ to denote the set of input Boolean variables. A Boolean vector $x \in \{0,1\}^n$ is a model of $f$ if $f(x) = 1$. Otherwise, $x$ 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 $\land$ (conjunction), $\lor$ (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 $\pi_x$ (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 $x \in \{0,1\}^n$, the corresponding canonical term is $t_x = \bigwedge_{i=1}^{n} x_i^{v_i} \land \pi_{\neg x_i}$, where $x_i^{v_i} = 1$ if $v_i$ is true for $x$. A term $t$ covers an assignment $x$ if $t \models t_x$. A satisfiable term $t$ is an implicat of $f$ if and only if $t \models f$ holds, and it is a prime implicat of $f$ if and only if $t$ is an implicat of $f$ and no proper subset of $t$ is an implicat of $f$. A non-valid clause $c$ is an implicat of $f$ if and only if $f \models c$ holds, and $c$ is a prime implicat of $f$ if and only if $c$ is an implicat of $f$ and no proper subset of $c$ is an implicat of $f$. A DNF formula is a conjunction 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 $\mathcal{DT}_n$, and the class of random forests over $X_n$ is denoted $\mathcal{RF}_n$.

**3 Improving Generality by Rewriting Instances**

Two families of contrastive explanations for instances from $X$ 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 $A$ and a tree-based binary classifier $f$ over $A$:

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

Accordingly, instances $x$ 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 $r_f(x)$ over the Boolean variables $C_f$ used by the predictor $f$. The rewrite function $r_f$ we consider is thus a mapping from $X$ to the terms over $C_f$.

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 $r_f(x)$ is logically connected to $x$ via a domain theory. Formally, for every numerical attribute $A_i$, one assumes an implicit First-Order Logic (FOL) theory capturing the semantics of $= \leq$ 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 catégorical 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 $(A_i, v_q)$. Instead of the FOL theory itself, when dealing with an instance $x \in X$, it is enough to consider the propositional grounding of the theory given $x$ and $x = (v_1, \ldots, v_n)$. Using the propositional grounding of the theory instead of the theory itself is more convenient from a computational perspective. For any numerical attribute $A_i \in A_{num}$, let $D_f(A_i) = \{v^1, \ldots, v_n\}$ be the set of values – ordered in ascending way (i.e., $v_1 < \cdots < v_n$) – about $A_i$ that can be found in the decision nodes of $f$. The corresponding propositional grounding is the formula

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

where

$$\Sigma_{num}(A_i, f) = \bigwedge_{j=1}^{n} ((\neg A_i \lor v_j) \Rightarrow (A_i \geq v_{j+1}))$$

and

$$c(v_i, D_f(A_i)) = (\neg A_i \lor v_i) \land (A_i \geq v_{i+1})$$

when $v_i \leq v_{i+1}$, and $c(v_i, D_f(A_i)) = (A_i \geq v_i)$ when $v_i < v_{i+1}$.

For any categorical attribute $A_i \in A_{cat}$, let $D_f(A_i) = \{v_1, \ldots, v_n\}$ be the set of values that can be found in the decision nodes of $f$. The corresponding propositional grounding is given by the formula

$$\Sigma_{cat}(A_i, f, x) = \bigwedge_{j=1}^{n} ((A_i = v_j) \Rightarrow (A_i = v_{j+1})) \land \Sigma_{cat}(A_i, f)$$

where

$$\Sigma_{cat}(A_i, f) = \Sigma_{num}(A_i, f, x) \land \Sigma_{cat}(A_i, f, x)$$

In $\Sigma_{num}(A_i, f, x)$ and $\Sigma_{cat}(A_i, f, x)$, $A_i = v_i$ and $A_i = v_j$ ($j \in [n]$) 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 $x$ (and more precisely of the value $v_i$ taken by $A_i$ in $x$) 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 $\bigwedge_{A_i \in A_{num}} \Sigma_{num}(A_i, f) \land \bigwedge_{A_i \in A_{cat}} \Sigma_{cat}(A_i, f)$. We are now in position to make more formal the notion of rewritten instance.

**Definition 2.** Let $x = (v_1, \ldots, v_k) \in X$ be an instance over $A = \{A_1, \ldots, A_\kappa\}$ where $(A_i = v_i) \in C$ (i $\in [k]$). Let $f$ be a tree-based classifier over $A$. The rewritten instance $t_r(x)$ over $C_f$ is given by $t_r(x)$ where $t_r(x)$ is the set of all literals over $C_f$ that are logical consequences of $t_e$ given $\Sigma(f, x) = \bigwedge_{A_i \in A_{num}} \Sigma_{num}(A_i, f, x) \land \bigwedge_{A_i \in A_{cat}} \Sigma_{cat}(A_i, f, x)$.

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

In the general case, $t_r(x)$ is not equivalent given $\Sigma(f, x)$ to $t_e$ but is strictly more general. Thus, on the running example, $t_r(x)$ captures not only Alice but the whole population of instances $x \in X$ with less than $20\%$ 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 $v'_i$ can be considered in the tree-based classifier $f$ for the same numerical attribute $A_i$ from $C$. Of course, it can be the case that an attribute $A_i$ from $C$ 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 $C$ 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 $C_f$.

Because Boolean conditions in $C_f$ are connected when they are issued from the same attribute $A_i \in A$, $\Sigma(f)$ must be taken into account to discard explanations that do not comply with $\Sigma(f)$, and are not legit as a consequence [47]. As a matter of illustration, consider Alice’s case again. $c = \{(A_1 = 30)\}$ is a subset-minimal contrastive explanation for $x$ given $\Sigma(f)$ in the sense of [24]. However, no instance from $X$ matches this representation in $C_f$ because $t_r(x (A_1 = 30)) = \{(A_1 = 30), (A_1 = 20), (A_2 = 1)\}$ conflicts with $\Sigma(f) = \{(A_1 = 20) \Rightarrow (A_1 = 30)\}$. 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 $\Sigma$, 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 [18].

**Definition 3.** [18] Let $X_n = \{x_1, \ldots, x_n\}$ be a set of Boolean variables. A constrained decision-function over $X_n$ is a pair $(f, \Sigma)$ where $f \in F_n$ and $\Sigma$ is a propositional formula over $X_n$. $\Sigma$ 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 $(f, \Sigma)$ be a constrained decision-function and $x \in \Sigma$ be an instance.

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

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

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

- An abductive explanation for $x$ given $(f, \Sigma)$ is a set $t \subseteq t_e$ such that $t \land \Sigma \models f$ (resp. $t \land \Sigma \models \neg f$).
- A subset-minimal abductive explanation for $x$ given $(f, \Sigma)$ is an abductive explanation $t$ for $x$ given $(f, \Sigma)$ such that no proper subset of $t$ is an abductive explanation for $x$ given $(f, \Sigma)$.
- A minimum-size abductive explanation for $x$ given $(f, \Sigma)$ is an abductive explanation $t$ for $x$ given $(f, \Sigma)$ such that no abductive explanation $t'$ for $x$ given $(f, \Sigma)$ such that $|t'| < |t|$ exists.

Subset-minimal abductive explanations (alias sufficient reasons) given a constrained decision-function have been investigated in [18] (see also [12] 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 [24] that still holds when a domain theory $\Sigma$ is taken into account [47], 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 [18] and [14], as well as the notion of universal literal quantification considered in [15]. Let us recall this notion. If $\ell$ is a literal over $x$, then the universal quantification of $\ell$ from $f$, noted $\forall x \cdot f$, is the formula $(\ell \lor (f[\ell])) \land (f[\ell])$. In this expression, $(f[\ell])$ denotes the conditioning of $f$ by $\ell$. If $x = (x)$ is a positive literal (resp. $\ell = \exists x$ 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, $\forall t \cup \{t_{k+1}\} : f$ denotes the formula $\forall t \cdot (\forall t_{k+1} \cdot f)$. Finally, $\forall x$ is a short for $\forall x$.

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 \((f, \Sigma)\) be a constrained decision-function and \(x \in [\Sigma]\) be an instance s.t. \(f(x) = 1\) (resp. \(f(x) = 0\)).

- The contrastive explanations for \(x\) given \((f, \Sigma)\) are the sets of literals \(c\) such that \(\forall \epsilon < \ell\ \epsilon\) is an implicate of \(\forall x : (\Sigma \Rightarrow f)\) (resp. \(\forall x : (\Sigma \Rightarrow \neg f)\)).
- The subset-minimal contrastive explanations for \(x\) given \((f, \Sigma)\) are the sets of literals \(c\) such that \(\forall \epsilon < \ell\ \epsilon\) is a prime implicate of \(\forall x : (\Sigma \Rightarrow f)\) (resp. \(\forall x : (\Sigma \Rightarrow \neg f)\)).
- The minimum-size contrastive explanations for \(x\) given \((f, \Sigma)\) are the sets of literals \(c\) such that \(\forall \epsilon < \ell\ \epsilon\) is a minimum-size prime implicate of \(\forall x : (\Sigma \Rightarrow f)\) (resp. \(\forall x : (\Sigma \Rightarrow \neg f)\)).

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

\[ t_{rf(x^0)} = \{(A_1 \geq 30), (A_2 \geq 20), (A_2 = 1)\}. \]

\(f\) is equivalent to \((A_1 \geq 30) \lor ((A_1 \geq 20) \land (A_2 = 1))\) and \(\Sigma = \Sigma(f) = (A_1 \geq 30) \lor (A_1 \geq 20)\).

Thus, \(\Sigma \Rightarrow \neg f\) is equivalent to \((A_1 \geq 20) \lor ((A_1 \geq 30) \land (A_2 = 1))\) and \(\forall \epsilon \geq 1\) \(f\) is equivalent to \((A_1 \geq 20) \lor (A_2 = 1)\).

This formula has two prime implicates:

- \((A_1 \geq 20) \lor (A_2 = 1)\)
- \((A_1 \geq 30) \land (A_2 = 1)\).

Accordingly, \(r_f(x^0)\) has two subset-minimal contrastive explanations given \((f, \Sigma)\), namely

\[ c_1 = \{(A_1 \geq 20), (A_1 \geq 30)\} \quad \text{and} \quad c_2 = \{(A_1 \geq 20), (A_2 = 1)\}. \]

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

\[ t_{rf(x^0)} = \{(A_1 \geq 30), (A_1 \geq 20), (A_2 = 1)\}, \]

\[ t_{rf(x^0)} = \{(A_1 \geq 30), (A_1 \geq 20), (A_2 = 1)\}. \]

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 an 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 \(\Sigma\) is valid – e.g., \(\Sigma = 1\)), an instance \(x\) 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 \(RF_n\), and \(x \in \{0, 1\}^n\) be an instance. The number of minimum-size contrastive explanations for \(x\) 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 \(x\) 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) \([4]\). In particular, when \(f(r_f(x)) = 0\), those characteristics are given by the literals implying \(\forall \epsilon \geq 1\) \(f\) on \(\epsilon\) depends on \([32]\). Thus, on Alice’s example, \((A_1 \geq 20)\) is the unique necessary characteristic of subset-minimal contrastive explanations for \(r_f(x^0)\), while all the characteristics in \(t_{rf(x^0)}\) 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 \(x \in [\Sigma]\) always has an abductive explanation given \((f, \Sigma)\) (indeed, \(t_a\) is such an abductive explanation), \(x\) does not always have a contrastive explanation given \((f, \Sigma)\). To be more precise:

**Proposition 3.** Let \((f, \Sigma)\) be a constrained decision-function and \(x \in [\Sigma]\) be an instance such that \(f(x) = 1\) (resp. \(f(x) = 0\)). \(x\) has a contrastive explanation given \((f, \Sigma)\) if and only if \(\neg f \land \Sigma\) is satisfiable. Deciding whether \(x\) has a contrastive explanation given \((f, \Sigma)\) is \(NP\)-complete. \(NP\)-hardness still holds when \(f\) is represented by a random forest from \(RF_n\) 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 \(t_a\) 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 \(RF_n\) and \(\Sigma = 1\) \([5]\)).

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

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

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

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

Subset-minimal and minimum-size abductive explanations are harder to recognize, even in the unconstrained case (i.e., when \(\Sigma\) is valid). Indeed, in the case of a random forest \(F \in RF_n\), deciding whether \(t \subseteq t_a\) is a subset-minimal abductive explanation for \(x\) given \(F\) has been shown DP-complete \([27]\) (the membership to DP extends to the general case when the classifier is any Boolean function \(f \in \mathcal{F}_n\)). In \([5]\), it has been shown that, given \(x \in \{0, 1\}^n\), \(F \in RF_n\) such that \(F(x) = 1\), and an integer \(k\), deciding whether there exists a minimum-size abductive explanation \(t\) for \(x\) given \(F\) such that \(|t| \leq k\) is \(\Sigma^2\)-complete. On this basis, one can show that deciding whether \(t\) is a minimum-size abductive explanation for \(x\)
given $F$ is $\Sigma_2^P$-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 [6], while the latter can be done in (deterministic) polynomial time (see e.g., [22]).

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 an instance $x$ given $(f, \Sigma)$. By definition, such contrastive explanations are less numerous than the contrastive explanations for $x$ given $(f, \Sigma)$ and they express smaller changes (in terms of the number of characteristics of $x$ to be modified).

A Partial MAXSAT characterization The following proposition shows how to derive a minimum-size contrastive explanation for an instance $x$ given a constrained decision-function $(f, \Sigma)$ using a Partial MAXSAT solver.

Proposition 7. Let $(f, \Sigma)$ be a constrained decision-function and $x \in [\Sigma]$ be an instance such that $f(x) = 1$. Let $(C_{soft}, C_{hard})$ be an instance of the Partial MAXSAT problem such that $C_{soft} = t_x$ and $C_{hard} = CNF(\Sigma \land \bar{f})$ where $CNF(\Sigma \land \bar{f})$ is a CNF encoding of $\Sigma \land \bar{f}$. Let $z^*$ be an optimal solution of $(C_{soft}, C_{hard})$. Then, $c = t_x \setminus t_{z^*}$ is a minimum-size contrastive explanation for $x$ given $(f, \Sigma)$ and we have $t_{z^*} = t_{z^*} \cap L_X$.

Let us recall that an instance of Partial MAXSAT consists of a pair $(C_{soft}, C_{hard})$ where $C_{soft}$ and $C_{hard}$ are (finite) sets of clauses. According to Proposition 7, finding a minimum-size contrastive explanation $c$ for $x$ given $f$ then mainly amounts to finding an assignment $z$ of the propositional variables involved in $C_{soft} \cup C_{hard}$ that maximizes the number of clauses in $C_{soft}$ that are satisfied, while satisfying all clauses in $C_{hard}$. Here $CNF(\Sigma \land \bar{f})$ denotes any CNF formula that is query-equivalent to $\Sigma \land \bar{f}$, 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 [46] or Plaisted/Greenbaum one [40] – 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 $C_{hard}$ the negation of the corresponding contrastive instance $t_{z^*}$ 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.

Interestingly, the Partial MAXSAT setting offers some flexibility that can be leveraged for generating valuable contrastive explanations, reflecting preferences of the explainer. For instance, additional hard constraints can be easily taken into account if one wants to ensure that the minimum-size contrastive explanation that is generated does not include any non-actionable characteristics. As a matter of illustration, suppose that Alice cannot reimbursed her previous loan, so that flipping the truth value of $(A_2 = 1)$ is not an option for her. If a characteristic $x_i$ (resp. $x_j$) of $x$ cannot be updated, then it is enough to consider $C_{hard} = CNF(\Sigma \land \bar{f} \land x_i)$ (resp. $C_{hard} = CNF(\Sigma \land \bar{f} \land x_j)$). In Alice’s case, $(A_2 = 1)$ would be added as a hard constraint. More generally, whenever the infeasibility of a contrastive explanation comes from a combination of features that is given by a propositional representation $\alpha$ over $X_n$, one may consider $C_{hard} = CNF(\Sigma \land \bar{f} \land \alpha)$. Thus, if it is not possible to change both the characteristics $x_i$ and $x_j$ of $x$, i.e., it is expected that the contrastive explanation to be derived must not satisfy $\alpha = \bar{f} \land \bar{x_i} \lor \bar{x_j}$, then considering $C_{hard} = CNF(\Sigma \land \bar{f} \land (x_i \lor x_j))$ is enough. Such actionability constraints is a way to implement dichotomous preferences over explanations [3].

More sophisticated preferences can also be taken into account. Thus, some characteristics can be grouped together by adding specific hard constraints. This can be useful when the corresponding Boolean attributes come from the same numerical attribute that must be counted only once if its value must be updated. Thus, in Alice’s case, the two Boolean attributes $(A_1 \geq 30)$ and $(A_1 \geq 20)$ of $C_1$ are issued from the same numerical attribute $A_1$ of $C$. Suppose that Alice is ready to consider how to increase her annual incomes or how to reimburse her previous loan, but not both of them at the same time. If such a preference is considered, $c_1 = \{(A_1 = 20), (A_1 = 30)\}$ must be preferred to $c_2 = \{(A_1 = 20), (A_2 = 1)\}$, since the latter refers to two attributes of $C$ $(A_1$ and $A_2$). To deal with this scenario, one can consider the following instance of Partial MAXSAT: $C_{hard} = CNF(\Sigma \land f \land (A_1 \lor (A_1 \geq 20)) \land (A_1 \lor (A_1 \geq 30)))$, where $A_1$ is a new Boolean variable associated with the numerical attribute with the same name, and $C_{soft} = \{(A_1, (A_1 = 20)), (A_1, (A_1 = 30))\}$ is used to group the two characteristics $(A_1 \geq 20)$ and $(A_1 \geq 30)$ of $r_f(x^*)$. By construction, $A_1$ is set to 1 in a solution if and only if the truth values of the Boolean variables $(A_1 \geq 20)$ and $(A_1 \geq 30)$ that occur in the solution are the same ones as in $t_{z^*}(\alpha)$, i.e., here they are set to 0. It can be easily checked that there is only one optimal solution for this instance of Partial MAXSAT, and it leads to $c_1 = \{(A_1 = 20), (A_1 \geq 30)\}$, as expected. Refined preferences can also be handled by switching from the Partial MAXSAT setting to the WEIGHTED PARTIAL MAXSAT setting. In such a framework, weights are added to soft constraints and an optimal solution maximizes the sum of the weights of the clauses in $C_{soft}$ that are satisfied. This leads to a utility function based on the characteristics of the input instance $x$ that indicates to which extent preserving those characteristics is important [3]. For instance, if Alice believes that increasing her annual incomes to at least 30k$ will be more difficult than both increasing her annual incomes to 20k$ and reimbursing her previous loan, Alice can express her preference by giving a weight of 3 to $(A_1 \geq 30)$, and a weight of 1 to $(A_1 \geq 20)$ and to $(A_2 = 1)$. That way, the unique contrastive explanation for $r_f(x^*)$ that is computed is $c_2 = \{(A_1 = 20), (A_2 = 1)\}$.

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 stan-
dard 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) [39]. 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 was used as Σ(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 that 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 x over 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 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 [35].

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 an Intel(R) XEON ES-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 A its is based on, the mean number of Boolean variables from C_F used in the random forests that have been generated, the mean number of attributes from 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 C_F in c, the number of attributes from A the literals of c are issued from, the percentage of literals in c relative to |C_F|, and finally the percentage of number of attributes from A the literals of c are issued, relative to |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 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 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 [36]). 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 x one started with. Changing the values of a few percentage (in average, 12.81%) of the attributes of A used in F is enough to change the way 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 [28, 38, 19] for recent references) using optimization techniques (e.g., MILP). Various distances / norms over 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 [33]. Approximately nearest contrastive instances with a preset degree of accuracy can also be considered [29]. 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 [45].

In our approach, instances are considered over C_F, thus described using Boolean attributes. In such a case, contrastive explanations correspond to sets of contrastive instances over C, hence they are typically more general than single contrastive instances. The minimum-size contrastive explanations for a rewritten instance r_F(x) are the (provably) nearest contrastive instances of r_F(x) over C_F w.r.t. Manhattan/Hamming distance (or, equivalently, l_1-norm). No commensurability assumptions about the scales of numerical attributes of A is needed in our approach, while the difficult task of identifying meaningful scaling factors has to be achieved when local distances over D_A are aggregated to define a distance over 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 [13], 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 \( X \) 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 \( x \) and different norms, including the \( \ell_0 \)-norm over \( C \), which corresponds to the \( \ell_0\)-norm over \( C_f \) as in our work. Domain theories for numerical attributes are considered implicitly encoded in the pointed out in [13], while our approach takes advantage of any explicit domain theory in QF format. Besides, the encoding used in [13] 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, [13] 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 \( x \) can be considered either as they are or are alternatively, as instances \( r_f(x) \) 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 \( (f, \Sigma) \) 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.

Aknowledgements

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 EXPERKCTATION (ANR-19-CHAIA-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


[16] Y. Freund and R.E. Schapire, ‘A decision-theoretic generalization of


Proofs

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

- The contrastive explanations for \(x\) given \((f, \Sigma)\) are the sets of literals occurring in the implicates of \(\forall x \cdot (\Sigma \Rightarrow f)\) (resp. \(\forall x \cdot (\Sigma \Rightarrow \overline{f})\)).
- The subset-minimal contrastive explanations for \(x\) given \((f, \Sigma)\) are the sets of literals occurring in the prime implicates of \(\forall x \cdot (\Sigma \Rightarrow f)\) (resp. \(\forall x \cdot (\Sigma \Rightarrow \overline{f})\)).
- The minimum-size contrastive explanations for \(x\) given \((f, \Sigma)\) are the sets of literals occurring in the minimum-size prime implicates of \(\forall x \cdot (\Sigma \Rightarrow f)\) (resp. \(\forall x \cdot (\Sigma \Rightarrow \overline{f})\)).

Proof. The proof of Proposition 1 from [18] shows that the abductive explanations (resp. subset-minimal abductive explanations) for \(x\) given \((f, \Sigma)\) are the sets of literals occurring in the implicates (resp. prime implicates) of \(f\) when \(f(x) = 1\). As a direct consequence, we also have that the abductive explanations (resp. subset-minimal abductive explanations) for \(x\) given \((f, \Sigma)\) are the sets of literals occurring in the implicates (resp. prime implicates) of \(\Sigma \Rightarrow \overline{f}\) such that \(t \subseteq t_w\) when \(f(x) = 0\). Then, we take advantage of the notion of universal literal quantification considered in [15] and use Proposition 11 from [14] to get that the sets of literals occurring in the implicates (resp. prime implicates) of \(\forall x \cdot (\Sigma \Rightarrow f)\) are the contrastive explanations (resp. subset-minimal contrastive explanations) for \(x\) given \((f, \Sigma)\) when \(f(x) = 1\), and that the sets of literals occurring in the implicates (resp. prime implicates) of \(\forall x \cdot (\Sigma \Rightarrow \overline{f})\) are the contrastive explanations (resp. subset-minimal contrastive explanations) for \(x\) given \((f, \Sigma)\) when \(f(x) = 0\). Finally, the previous result about subset-minimal contrastive explanations extend to minimum-size contrastive explanations, given that the minimum-size contrastive explanations for \(x\) given \((f, \Sigma)\) are the subset-minimal contrastive explanations for \(x\) given \((f, \Sigma)\) that of minimum size.

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

Proof. Let \(k = \lceil \frac{n}{2} \rceil\). Consider the DNF formula \(f = \bigwedge_{i=0}^{k-1} (x_{2i+1} \land x_{2i+2})\) and the instance \(x \in \{0,1\}^n\) such that \(x_i = 1\) for each \(i \in [n]\). We have \(\forall x \cdot f \equiv f\). The subset-minimal contrastive explanations for \(x\) given \((f, 1)\) are the sets of literals occurring in the prime implicates of \(\forall x \cdot f\), thus the sets of literals occurring in the prime implicates of \(f\). They all have the same size \((k)\), hence they are all minimal-size contrastive explanations for \(x\) given \((f, 1)\). Consider now a random forest \(F\) from \(\mathbb{RF}_n\) equivalent to \(f\) and containing \(m = 2k - 1\) trees (see Proposition 2 from [5] for the generation of \(F\) from \(f\)). The fact that \(f\) has \(2^k\) prime implicates completes the proof.

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

Proof. Suppose that \(x \in [\Sigma]\) is such that \(f(x) = 1\). If \(\overline{f} \land \Sigma\) is unsatisfiable then there is no model \(x' \in \Sigma\) that is a model of \(\overline{f}\), hence \(x\) does not have any contrastive explanation given \((f, \Sigma)\). Similarly, if \(x \in [\Sigma]\) is such that \(f(x) = 0\) and \(f \land \Sigma\) is unsatisfiable then there is no model \(x' \in \Sigma\) that is a model of \(f\), hence \(x\) does not have any contrastive explanation given \((f, \Sigma)\). Then the complexity results can be derived as follows:

- Membership to \(\mathsf{NP}\): The following nondeterministic algorithm runs in time polynomial in the input size \((x\) and a representation of \(f))\) that \(x\) and \(f(x) \neq f(\Sigma)\).
- \(\mathsf{NP}\)-hardness: we prove that the restriction of the decision problem when \(f\) is a random forest \(F\) from \(\mathbb{RF}_n\) and \(\Sigma = 1\) is \(\mathsf{NP}\)-hard by reduction from the satisfiability problem for \(\mathbb{CNF}\) formulae. Let \(\alpha = c_1 \land \cdots \land c_m\) be a \(\mathbb{CNF}\) formula over \(X_n\). Let \(x\) be any instance from \(\{0,1\}^n\) such that \(\{T \in C \} \subseteq t_w\). We associate with \(\alpha\) in polynomial time the pair \((x, F)\), where \(F\) is a random forest from \(\mathbb{RF}_n\) equivalent to \(\overline{\alpha}\) (see Proposition 2 from [5] for the generation of \(F\)). We have \(f(x) = 1\) since \(\alpha(x) = 0\) by construction of \(x\). Since deciding whether \(x\) has a contrastive explanation given \((F, 1)\) amounts to deciding whether \(\overline{F}\) is satisfiable and since \(\overline{F}\) is equivalent to \(\alpha\), the conclusion follows.

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

Proof. By definition, \(c\) is a contrastive explanation for \(x\) given \((f, \Sigma)\) if and only if \(x_c \in [\Sigma]\) and \(f(x_c) \neq f(x)\). Both tests can be achieved in polynomial time since \(x_c\) is an interpretation over \(X_n\) and \(f\) and \(\Sigma\) are built upon \(X_n\).

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

Proof.

- Membership to \(\mathsf{coNP}\): we first check that \(c\) is a contrastive explanation for \(x\) given \((f, \Sigma)\). This is done in polynomial time (see Proposition 4). Now, \(c\) is not a sub-minimal contrastive explanation for \(x\) given \((f, \Sigma)\) if and only if there exists a proper subset \(c'\) of \(c\) that is a contrastive explanation for \(x\) given \((f, \Sigma)\). Deciding whether such a \(c'\) exists is in \(\mathsf{NP}\): it is enough to guess \(c' \subset c\) and to test in deterministic polynomial time that \(c'\) is a contrastive explanation for \(x\) given \((f, \Sigma)\).
- \(\mathsf{coNP}\)-hardness: we prove that the restriction of the decision problem when \(f\) is a random forest \(F\) from \(\mathbb{RF}_n\) and \(\Sigma = 1\) is \(\mathsf{coNP}\)-hard by reduction from the minimal model checking problem for \(\mathbb{CNF}\) formulae, which is \(\mathsf{coNP}\)-complete [10]. The latter problem is as follows:

- Input: an instance \(x \in \{0,1\}^n\) and a \(\mathbb{CNF}\) formula \(\alpha = \bigwedge_{i=1}^m c_i\) over \(X_n\).
- Question: Is \(t_w\) a sub-minimal model of \(\alpha\), i.e., a model of \(\alpha\) such that the set of positive literals in \(t_w\) is minimal w.r.t. set-inclusion?
We first prove the following lemma:

**Proposition 6.** Let \( x \in \{0,1\}^n \) and \( \alpha = \bigwedge_{i=1}^m c_i \) be a CNF formula over \( X_n \). Let \( z \) be the instance of \( \{0,1\}^{n+1} \) such that \( z_i = 0 \) for \( i \in [n+1] \). Let \( c = \{\pi_T : x_i = 1, i \in [n] \} \cup \{\pi_{x_{n+1}}\} \). \( c \) is a contrastive explanation for \( z \) given \( \langle \pi \lor x_{n+1}, 1 \rangle \) if and only if \( t_x \) is a model of \( \alpha \). Furthermore, \( z_\alpha \) coincides with \( x \) over \( X_n \).

**Proof.** We have that \( t_x \) is a model of \( \pi \lor x_{n+1} \) since \( \pi_{x_{n+1}} \in t_x \). We also have \( c \subseteq t_x \) since \( c \) contains only negative literals. \( c \) is a contrastive explanation for \( z \) given \( \langle \pi \lor x_{n+1}, 1 \rangle \) if and only if \( z_\alpha \) is a model of \( \alpha \land x_{n+1} \). Since \( c = \{\pi_T : x_i = 1, i \in [n] \} \cup \{\pi_{x_{n+1}}\} \), \( z_\alpha \) coincides with \( x \) over \( X_n \). Since, by construction, \( x_{n+1} \in t_x \), \( z_\alpha \) is a model of \( \alpha \land x_{n+1} \) if and only if \( z_\alpha \) is a model of \( \alpha \) if and only if \( t_x \) is a model of \( \alpha \).

Then the reduction from the minimal model checking problem is as follows: to any input \( (x, \alpha) \) of the minimal model checking problem we associate in polynomial time the following triple \( (F, \pi, c) \):

\[
F = \{T(\pi_{x \lor x_{n+1}}, \ldots, T(\pi_m \lor x_{n+1}), \mathbb{T}, \ldots, \mathbb{T})_{m-1}\}
\]

is a random forest over \( X_n \) containing \( 2m - 1 \) trees. Each \( T(\pi_{x \lor x_{n+1}}) \) (where \( x \) is a decision tree over \( X_n \)) is equivalent to the formula \( \pi \lor x_{n+1} \) (this tree can be generated in time linear in the size of \( c \)). By construction, \( F \) is equivalent to \( \pi \lor x_{n+1} \), so that \( F \) is equivalent to \( \alpha \land x_{n+1} \). Finally, take \( x \) and \( z_\alpha \) as given in Lemma 1. From Lemma 1, since \( z_\alpha \) coincides with \( x \) over \( X_n \) and since every model of \( F \) must set \( x_{n+1} \) to 1 as \( x \) does it, \( t_x \) is a minimal model of \( \alpha \) if and only if \( c \) is a minimal model of \( \alpha \).

We first prove that the minimum-cardinality model checking problem for CNF formulæ is coNP-hard. The reduction is from UNSAT: starting with a CNF formula \( \beta = \bigwedge_{i=1}^m c_i \) over \( X_n \), let us associate in polynomial time the CNF formula \( \alpha = \bigwedge_{i=1}^{2(n+1)} \bigwedge_{i=1}^m (x_i \lor c_i) \) over \( X_{2n+1} \) and the instance \( x = \{0,1\}^{2(n+1)} \) that every \( x_i \) (for \( i \in [n] \)) to 0 and every \( x_i \) (for \( i \in [n+1, 2n+1] \)) to 1. \( \alpha \) contains \( n+1 \) \( m \)-clauses and is of size \( \mathcal{O}(m^2) \). If \( \beta \) is unsatisfiable, then \( \alpha \) is equivalent to \( \bigwedge_{i=1}^{2(n+1)} x_i \). Hence \( t_x \) is the sole minimum-cardinality model of \( \alpha \) (it contains \( n+1 \) variables set to 1). If \( \beta \) is satisfiable, then it has a model over \( X_n \), thus it also has a minimum-cardinality model over \( X_n \), and this model contains at most \( n \) variables set to 1. Then the extension over \( X_{2n+1} \) of this minimum-cardinality model obtained by setting every \( x_i \) (for \( i \in [n+1, 2n+1] \)) to 0 is a minimum-cardinality model of \( \beta \). Since this model contains at most \( n \) variables set to 1, \( t_x \) is not a minimum-cardinality model of \( \alpha \). This concludes the coNP-hardness proof for the problem of checking a minimum-cardinality model for CNF formulæ.

Then we reduce the minimum-cardinality model checking problem to the problem of deciding whether \( c \) is a minimum-size contrastive explanation for \( x \) given \( F(1) \) where \( F \in \mathbb{R}_n \) is a random forest. To any input \( (x, \Sigma) \) of the minimum-cardinality model checking problem we associate in polynomial time the triple \( (F, \Sigma, c) \) as given in the proof of Proposition 5. We use Lemma 1 to conclude that \( t_x \) is a minimum-cardinality model of \( \Sigma \) if and only if \( c \) is a minimum-size contrastive explanation for \( x \) given \( F(1) \).

**Proposition 7.** Let \( (f, \Sigma) \) be a constrained decision-function and \( x \in \Sigma \) be an instance such that \( f(x) = 1 \). Let \( (C_{soft}, C_{hard}) \) be an instance of the PARTIAL MAXSAT problem such that \( C_{soft} = t_x \) and \( C_{hard} = CNF(\Sigma \land \overline{f}) \) where \( CNF(\Sigma \land \overline{f}) \) is a CNF encoding of \( \Sigma \land \overline{f} \). Let \( z^* \) be an optimal solution of \( C_{soft}, C_{hard} \). Then, \( c = t_x \setminus t_{x^*} \) is a minimum-size contrastive explanation for \( x \) given \( (f, \Sigma) \) and we have \( t_{x^0} = t_{x^*} \cap L_{X_n} \).

**Proof.** Let \( z^* \) be any optimal solution of \( (C_{soft}, C_{hard}) \). On the one hand, \( z^* \) is a model of \( C_{hard} \). Since \( \Sigma \land \overline{f} \) is equivalent to \( \Sigma \land \overline{f} \), \( t_{x^0} \) is a model of \( \Sigma \land \overline{f} \). Now, since \( z^* \) is an optimal solution of \( (C_{soft}, C_{hard}) \), \( z^* \) satisfies a maximal number of soft clauses from \( C_{soft} \). Since those soft clauses are precisely the literals occurring in \( t_{x^0} \), the set of literals \( c = t_x \setminus t_{x^*} \) is a subset of literals of \( t_{x^0} \) of minimum-size such that \( t_{x^0} = t_x \cap L_{X_n} \) is a model of \( \Sigma \land \overline{f} \). Stated otherwise, \( c \) is a minimum-size contrastive explanation for \( x \) given \( (f, \Sigma) \).

**Proposition 8.** Let \( F \in \mathbb{R}_n \) such that \( F(x) = 1 \) for \( x \in \{0,1\}^n \), and \( t \subseteq t_x \). Then, \( c \subseteq t_x \) is a minimum-size contrastive explanation for \( x \) given \( F \) is \( \Pi_2 \)-complete.

**Proof.**

- Membership to \( \Pi_2 \): we show that the problem belongs to \( \Pi_2 \) in the general case when the classifier \( f \) is a Boolean function \( f \in \mathbb{F}_n \). To get the result, we prove that the complementary problem belongs to \( \Sigma_2^p \). Then, in order to decide whether \( t \) is not a minimum-size contrastive explanation for \( x \) given \( f \), it is enough to first test whether \( t \) is an explanation for \( x \) given \( f \) using one call to an \( \mathcal{NP} \) oracle; if \( t \) is not such an explanation, then it is not a
minimum-size abductive explanation for \( x \) given \( f \); in the remaining case, it is enough to guess \( t' \subseteq t_x \), to check that \( |t'| < |t| \), and finally to check (using one call to an NP oracle) that \( t' \) is an abductive explanation for \( x \) given \( f \).

\( \Pi^2 \_p \)-hardness: let us associate in polynomial time with \( \langle x \in \{0, 1\}^n, F = \{T_1, \ldots, T_m\} \in \mathbb{RF}, k \leq n \rangle \) such that \( F(x) = 1 \) the triple \( \langle x', F', t \rangle \) where \( x' \in \{0, 1\}^{n+k+1} \) coincides with \( x \) on its first \( n \) coordinates and is such that \( x'_j = 1 \) for each \( j \in [n+1, n+k+1] \), \( F' = \{T'_1, \ldots, T'_m\} \in \mathbb{RF}^{n+k+1} \) is such that \( T'_i \) (\( i \in [m] \)) is a decision tree equivalent to \( T_i \lor \bigwedge_{j=n+1}^{n+k+1} x_j \).

Clearly, each decision tree \( T'_i \) (\( i \in [m] \)) can be generated in time \( O(k \cdot |T_i|) \) given that the term \( \bigwedge_{j=n+1}^{n+k+1} x_j \) can be represented by a decision tree containing \( k \) decision nodes and that a decision tree representing the disjunction of two decision trees can be computed in time in the product of the sizes of the two trees (see [31]). Since \( k \leq n \), \( |F'| \) is polynomial in \( |x| \cdot |F| \). By construction, \( F' \) is equivalent to \( F \lor \bigwedge_{j=n+1}^{n+k+1} x_j \) so that \( F'(x') = 1 \). Finally, let \( t = \bigwedge_{j=n+1}^{n+k+1} x_j \). By construction, \( t \) is an implicant of \( F' \) such that \( t \subseteq t_x \). \( t \) contains \( k+1 \) characteristics. \( t \) is a prime implicant of \( F' \) unless \( F \) is valid (in which case \( F' \) is valid as well). More precisely, if \( F \) is valid, then \( \top \) is the unique prime implicant of \( F' \), else the prime implicants of \( F' \) are the prime implicants of \( F \), plus \( t \). So if \( F \) has a minimum-size abductive explanation for \( x \) given \( F \) containing at most \( k \) characteristics, then this explanation is also a minimum-size abductive explanation for \( x' \) given \( F' \), showing that \( t \) is not a minimum-size abductive explanation for \( x' \) given \( F' \). In the remaining case, every minimum-size abductive explanation for \( x \) given \( F \) contains at least \( k+1 \) characteristics (hence \( F \) is not valid). This shows that \( t \) is a minimum-size abductive explanation for \( x' \) given \( F' \), and this completes the proof.