Impact of Weight Functions on Preferred Abductive Explanations for Decision Trees

Louenas Bounia[1,2,*,†], Matthieu Goliot[2,†] and Anasse Chafik[1,2,†]

[1]Centre de Recherche en Informatique de Lens (CRIL), Université d'Artois & CNRS, France
[2]Université d'Artois, Lens, France

Abstract

In this article, our main objective is to address the issue of diversity in abductive explanations for decision trees by studying the impact of different weight functions on preferred abductive explanations. We acknowledge that users may have specific preferences regarding the explanations they prefer to receive. Therefore, we propose several criteria to obtain high-quality subsets of abductive explanations that take into account these preferences. These criteria are defined by the users themselves by assigning weights to different preference criteria. To evaluate the impact of these preference criteria on abductive explanations and the relationships between the obtained subsets, we propose an approach based on SAT encoding. This allows us to enumerate more easily the different subsets of abductive explanations that meet the user-defined preference criteria. Additionally, we use measures based on the distance between two sets of explanations to assess the correlation between user preferences and the extent to which result sets differ from each other for different preferences. In summary, this study represents the first step towards providing a framework for selecting abductive explanations that cater to users' preferences in a diverse and high-quality manner. We aim to instill the necessary confidence in users to utilize these explanations in their decision-making process by offering explanations tailored to their individual preferences.

Keywords

Explainable AI, Diversity of explanations, Decision trees, Weight functions

1. Introduction

Explaining Machine Learning (ML) models is an important challenge that has been a subject of study of AI in recent years (see, for example, [1, 2, 3, 4]. In this article, we focus on *abductive explanations* for binary decision tree models [5]. Abductive explanations aim to clarify *why* a classifier classifies an instance as positive or negative. In contrast, contrastive explanations aim to explain why the instance was not classified as expected (thus addressing the question "why not the other classification?"). Several types of abductive explanations exist depending on the used classifier. These include the *direct reason* [6], the *prime implicant* [7], also known as the *sufficient reason* [8]. The quality of an explanation relies not only on the reason itself but often depends on the person being explained to and the domain involved.

In this article, we focus on the diversity of abductive explanations, a crucial aspect when it comes to user-guided explanations. When a user requests an explanation for the classification of an example by a machine learning model, they may have specific preferences regarding the form or content of that explanation. For instance, some users prefer concise and succinct explanations, while others prioritize more detailed and comprehensive explanations. Our study primarily

ICCBR XCBR'23: Workshop on Case-Based Reasoning for the Explanation of Intelligent Systems at ICCBR2023, July 17 – 20, 2023, Aberdeen, Scotland

*Corresponding author.
†These authors contributed equally.
bounia@cril.fr (L. Bounia); matthieu_goliot@ens.univ-artois.fr (M. Goliot); chafik@cril.fr (A. Chafik)

centers on preferred abductive reasons, which are considered the most anticipated explanations by users. We have chosen to investigate the diversity of preferred explanations within the context of decision trees, which are widely used machine learning models. Diversity, in this context, can be perceived as a mean to account for different priorities among users. In other words, the objective of this study is to consider user preferences, specially when they vary from one another.

We first propose a SAT encoding based on the encoding proposed by Jabbour et al. [9] to enumerate the preferred sufficient reasons. Several weight functions based on XAI methods known in the literature have been considered to calculate the preferred reasons based on the weights provided by these functions. These weight functions allow us to calculate the preferred sufficient reasons for a given method (or a given user) using a gradual preference model expressed by weights. Finally, we evaluate the impact of different weight functions on the preferred sufficient reasons for a given decision tree, by first counting their number and then calculating the distance between two sets of preferred explanations. This measure allows us to quantify the gap between two subsets of explanations and thus measure the impact of user preference diversity on the produced explanations.

2. Decision Trees and Abductive Explanations

2.1. Preliminaries

For an integer n, let $[n]$ be the set $\{1, \ldots, n\}$. We denote $\mathcal{F}_n$ as the class of all Boolean functions from $\{0,1\}^n$ to $\{0,1\}$, and we use $X_n = \{x_1, \ldots, x_n\}$ to represent the set of Boolean input variables. Any assignment $x \in \{0,1\}^n$ is called an *instance*. If $f(x) = 1$ for $f \in \mathcal{F}_n$, then x is called a *model* of f. x is a *positive instance* if $f(x) = 1$, and a *negative instance* if $f(x) = 0$.

We refer to f as a *propositional formula* when it is described using the Boolean connectors $\wedge$ (conjunction), $\vee$ (disjunction), $\neg$ (negation), as well as the Boolean constants 1 (true) and 0 (false). Other connectors, such as implication $\rightarrow$, may also be considered. As usual, a *literal* ℓ is a variable x_i (a positive literal) or its negation $\neg x_i$, also denoted $\overline{x}_i$ (a negative literal). x_i and $\overline{x}_i$ are complementary literals. A positive literal x_i is associated with a positive feature (i.e., x_i is assigned to 1), while a negative literal $\overline{x}_i$ is associated with a negative feature.

A *term* t is a conjunction of literals, and a *clause* c is a disjunction of literals. $Lit(f)$ denotes the set of all literals in f. A DNF (Disjunctive Normal Form) formula is a disjunction of terms, and a CNF (Conjunctive Normal Form) formula is a conjunction of clauses. The set of variables appearing in a formula f is denoted by $Var(f)$. A formula f is *consistent* if and only if it has a model. A CNF formula is *monotone* when each literal of a given variable in the formula has the same polarity (i.e., each time a literal appears in the formula, the complementary literal does not appear in the formula). A formula f_1 *implies* a formula f_2, denoted $f_1 \models f_2$, if and only if every model of f_1 is a model of f_2. Two formulas f_1 and f_2 are *equivalent*, denoted $f_1 \equiv f_2$, if and only if they have the same models. Given an assignment $z \in \{0,1\}^n$, the corresponding term is defined as:

$$t_z = \bigwedge_{i=1}^{n} x_i^{z_i} \text{ où } x_i^0 = \overline{x}_i \text{ et } x_i^1 = x_i$$

A term t *covers* an assignment z if $t \subseteq t_z$. An *implicant* of a Boolean function f is a term that implies f. A *prime implicant* of f is an implicant t of f such that no proper subset of t is an implicant of f. Conversely, an *implicant* of a Boolean function f is a clause that is implied by f,

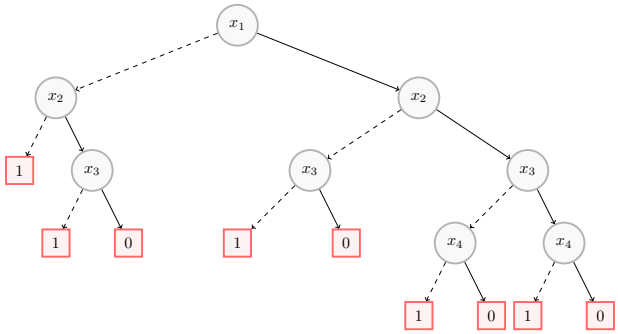

Figure 1: A decision tree T for classifies the allocation of a *bank loan*.

and a *prime implicant* of f is an implicant c of f such that no proper subset of c is an implicant of f.

Definition 1 (Boolean decision tree)**.** *A Boolean decision tree over X_n is a binary decision tree, where each internal node is labeled with one of the n Boolean input variables, and each leaf is labeled with either 0 or 1. Each variable appears at most once along any path from the root to a leaf. The value $T(\boldsymbol{x}) \in \{0,1\}$ of T for the input instance $\boldsymbol{x}$ is determined by the label of the leaf reached from the root as follows: at each node, we follow the left or right child depending on whether the input value of the corresponding variable is 0 or 1. The size of T (denoted $|T|$) is the number of nodes.*

The class of decision trees over X_n is denoted $\mathtt{DT}_n$. It is well-known that any tree $T \in \mathtt{DT}_n$ can be transformed into an equivalent disjunction of terms in linear time, denoted $\mathtt{DNF}(T)$, where each term corresponds to a path from the root to a leaf labeled 1. Similarly, T can be transformed in linear time into a conjunction of clauses, denoted $\mathtt{CNF}(T)$ [10], where each clause is the negation of a term corresponding to a path from the root to a leaf labeled 0.

The tree shown in Figure 1 will be used as an running example in the rest of the paper.

Example 1. *The decision tree in Figure 1 classifies* bank loans *using the following attributes: x_1: "does not have a permanent contract", x_2: "is over 50 years old", x_3: "has annual income below 35K" and x_4: "has not repaid a previous loan".*

2.2. Abductive explanations

We consider the concept of *abductive explanation*. Formally, for $f \in F_n$ and $\mathbf{x} \in \{0,1\}^n$, an *abductive explanation* (*reasons*) of $\mathbf{x}$ given f is an implicant t of f (or of $\neg f$ in the case where $f(\mathbf{x}) = 0$) that covers $\mathbf{x}$. There always exists an abductive explanation t of $\mathbf{x}$ given f because $t = t_{\mathbf{x}}$ is such a trivial explanation. Therefore, in the remainder of this section, we will focus on more concise forms of abductive explanation.

Direct reasons [10, 6] are abductive explanations specific to decision trees and random forests (see [11]). Other abductive explanations exist that are not specific to a particular classifier, such as *sufficient reasons* [8]. In the following, we will define sufficient reasons.

Definition 2 (Sufficient reason)**.** *Let $f \in \mathcal{F}_n$ and $\boldsymbol{x} \in \{0,1\}^n$ such that $f(\boldsymbol{x}) = 1$ (resp. $f(\boldsymbol{x}) = 0$). A* sufficient reason *for $\boldsymbol{x}$ given f is a prime implicant t of f (resp. $\neg f$) that covers $\boldsymbol{x}$. $sr(\boldsymbol{x}, f)$ denotes the set of all sufficient reasons for $\boldsymbol{x}$ given f.*

A *sufficient reason* [8] (or PI-explanation [7]) for an instance $\boldsymbol{x}$ given a Boolean function f is a subset t of $\boldsymbol{x}$ that is minimal with respect to set inclusion, and such that any instance $\boldsymbol{x'}$ that

shares the set t is classified by f as $\boldsymbol{x}$. Thus, when t covers $\boldsymbol{x}$, when $f(\boldsymbol{x}) = 1$, t is a sufficient reason for $\boldsymbol{x}$ given f if and only if t is a prime implicant of f, and when $f(\boldsymbol{x}) = 0$, t is a sufficient reason for $\boldsymbol{x}$ given f if and only if t is a prime implicant of $\neg f$. Sufficient reasons do not contain any redundant attributes. We refer to a *minimal-size sufficient reason* for $\boldsymbol{x}$ given f as a sufficient reason for $\boldsymbol{x}$ given f that contains the minimum number of literals.

Example 2. *Going back to Example 1, we can observe that $T(\boldsymbol{x}) = 0$ (Bank loan rejected.) for the instance $\boldsymbol{x} = (1, 1, 1, 1)$. The direct reason for $\boldsymbol{x}$ is $t_{\boldsymbol{x}}^T = x_1 \wedge x_2 \wedge x_3 \wedge x_4$, $x_1 \wedge x_2 \wedge x_4$, $x_1 \wedge x_3 \wedge x_4$ **and** $x_2 \wedge x_3 \wedge x_4$ are the sufficient reasons for $\boldsymbol{x}$ given T. They are also the only minimal-size sufficient reasons for $\boldsymbol{x}$ given T.*

3. Computing All Abductives Explanations

The number of sufficient reasons in an instance may be exponential [10]. In the following, we remind that even for the restricted class of decision trees with logarithmic depth, an instance $\boldsymbol{x}$ can have an exponential number of sufficient reasons. By definition, the number of minimal sufficient reasons for $\boldsymbol{x}$ cannot be greater than the number of its sufficient reasons. However, restricting ourselves to minimal sufficient reasons does not guarantee a significant reduction to their number [12, 10] because an instance can have an exponential number of minimal sufficient reasons. We shall recall a proposition that confirms the exponential nature of the number of minimal sufficient reasons which was proposed by Audemard et al. [10].

Proposition 1. *For any $n \in \mathbb{N}$ such that n is odd, there exists a decision tree $T \in \mathtt{DT}_n$ with depth $\frac{n+1}{2}$, containing $2n + 1$ nodes, and an instance $\boldsymbol{x} \in \{0, 1\}^n$ such that the number of minimum-size sufficient reasons for $\boldsymbol{x}$ given T is equal to $2^{\sqrt{n-1}}$.*

3.1. Compute all minimum-size sufficient reasons.

In order to synthesize the set of sufficient reasons, we first focus on the minimum-size sufficient reasons. Although the set of minimum-size sufficient reasons for an instance given a decision tree can be exponential, this number cannot exceed the total number of sufficient reasons, and in practice, it can be significantly smaller. However, unlike sufficient reasons, which can be generated in polynomial time [10, 12], computing the minimum-size reasons is not an easy task.

Proposition 2. *Let $T \in \mathtt{DT}_n$ and $\boldsymbol{x} \in \{0, 1\}^n$. Computing a minimum-size sufficient reason for $\boldsymbol{x}$ given T is* NP*-hard.*

Despite this result of intractability in the general case, computing a set of minimum-size sufficient reasons is possible in many practical cases. For this purpose, we rely on recent advancements in combinatorial optimization related to SAT.

First, let us recall that the Partial MaxSAT problem consists of a pair $(C_{\text{soft}}, C_{\text{hard}})$, where C_{soft} and C_{hard} are (finite) sets of clauses. The objective is to determine, if it exists, an assignment of variables that maximizes the number of satisfied clauses from C_{soft}, while satisfying all clauses from C_{hard}. We can utilize a Partial MaxSAT solver to compute minimal-size sufficient reasons:

Proposition 3. *Let T decision trees in $\mathtt{DT}_n$ and $\boldsymbol{x} \in \{0, 1\}^n$ an instance such that $T(\boldsymbol{x}) = 1$. Let $(C_{\text{soft}}, C_{\text{hard}})$ instance of Partial MaxSAT problem such that :*

$$C_{\text{soft}} = \{\overline{x_i} : x_i \in t_{\boldsymbol{x}}\} \cup \{x_i : \overline{x}_i \in t_{\boldsymbol{x}}\}$$

and

$$C_{\text{hard}} = \{c \cap t_{\boldsymbol{x}} : c \in \mathtt{CNF}(T)\}.$$

The intersection of $t_{\boldsymbol{x}}$ with $t_{\boldsymbol{x}^*}$, where $\boldsymbol{x}^*$ is an optimal solution for $(C_{\text{hard}}, C_{\text{soft}})$, is a minimal-size sufficient reason for $\boldsymbol{x}$ given T.

A Partial MaxSAT solver can also be used to compute a predefined number of minimal-size sufficient reasons. The process involves generating an initial reason t, adding the negation of t ($\neg t$) to C_{hard}, and including a cardinality constraint to ensure that the subsequent computed reasons have the same size as t. This process is repeated until the desired number of reasons is reached or no solution exists. Calculating a single explanation is often insufficient to fully understand the behavior of a classifier. On the other hand, providing millions of explanations would not be practical for the user. Reasons can vary greatly from one another, and the quality of a reason also depends on the person to whom it is explained. The authors of the article [13] propose leveraging user preferences to select the most relevant reasons and thus reduce their number. This restricted set of explanations has two advantages: it aligns as closely as possible with the user's preferences and can drastically reduce the overall number of explanations. However, it is important to note that even two experts on the same field may have different preferences. In our work, we focus on the impact of different weighting functions on the set of preferred sufficient reasons given a decision tree T, in order to better understand the diversity of abductive explanations.

4. Preferred abductive explanations

One rational way to address this question is to focus on a subset of explanations, referred to as the *preferred* ones [13]. Defining what makes an explanation "preferred" or "good enough" is challenging in general, and there is no consensus on this matter, as seen in [14]. Preferred explanations can be either the complete set of abductive explanations [15] or subsets thereof, particularly those containing only sufficient reasons. Although the notion of preferred reasons makes sense for any Boolean classifier, our results are specific to decision trees since they concern **sufficient reasons**. The authors of the paper [13] have defined several preference models, and in the following, we focus on one of them: *Maximum-Weight Explanations*.

4.1. Maximum-Weight Explanations

A model of preference relation on a combinatorial domain is by using a *utility function* (or cost function). In our context, this involves assigning a utility value (weight) to each feature. This approach leads to a total preorder on explanations, where the best explanations are those with the highest weight.

The idea behind a utility function is to measure the importance of each feature in the explanation. For example, one can assign a weight to each feature corresponding to its usefulness or relevance to the considered problem. The larger the utility value of a feature, the more important it is in the explanation. By associating a utility value with each feature, one can calculate an overall utility value for each explanation by summing the utility values of its features. This allows ranking explanations based on their utility value and determining the best explanations, those with the highest utility value. The advantage of this approach is that it allows for more complex preferences to be taken into account than simply ranking features in order of importance. Indeed, each user may have different preferences, and a personalized utility function allows for these preferences to be modeled more finely.

In the general case, computing a maximum-weight sufficient reason is NP-hard in the broad sense. This follows from the fact that a minimum-size sufficient reason t for a given instance of a decision tree is a minima-weight preferred reason t for a given instance and decision tree with

a weight mapping w_1 such that for each $i \in [n]$, $w_1(x_i) = 1$. Computing a maximum-weight sufficient reason t for a given instance of a decision tree is NP-hard [11, 16]. Nevertheless, the approach presented in [12] can be generalized to compute minimum-size sufficient reasons for the case of maximum-weight sufficient reasons. This amounts to solving an instance of the WEIGHTED PARTIAL MAXSAT problem.

Definition 3. *Let $T \in \mathtt{DT}_n$. Let $w : X_n \to \mathbb{N}^*$ a weight vector associated with each feature. A* maximum-weight reason *for $\boldsymbol{x}$ given T et w is a term t for $\boldsymbol{x}$ and T that maximize $\Sigma_{x \in Var(t)} w(x)$.*

Proposition 4. *Let $T \in \mathtt{DT}_n$ and an instance $\boldsymbol{x} \in \{0,1\}^n$ suche that $T(\boldsymbol{x}) = 1$. Let $w : X_n \to \mathbb{N}^*$ weights application. Maximum-weight sufficient reason for $\boldsymbol{x}$ given T et w is given by $t_{\boldsymbol{x}} \cap t_{\boldsymbol{v}^*}$, where $\boldsymbol{v}^*$ is the solution of $(C_{\text{soft}}, C_{\text{hard}})$ of WEIGHTED PARTIAL MAXSAT problem such that :*

$$C_{\text{soft}} = \{(x_i, w(x_i)) : x_i \in t_{\boldsymbol{x}}\} \cup \{(\overline{x}_i, w(x_i)) : \overline{x}_i \in t_{\boldsymbol{x}}\}$$
$$C_{\text{hard}} = \{(c[\boldsymbol{x}], \infty) : c \in \mathtt{CNF}_s(T)\}$$

where : C_{hard} : is the CNF *encoding proposed by [9] of the* CNF *encoding of decision tree T*

In the following, we will refer to "maximum-weight sufficient reason" as the explanation with the highest weight and "preferred sufficient reason" as the explanation preferred.

Remark. We would like to clarify that the encoding proposed in this article (Proposition 3) is different from the one proposed by the authors in [13], even though both are based on MaxSAT. The aim of the encoding in [13] is to minimize the sum of weights to obtain preferred reasons, while our approach aims to maximize it. Another major difference is the exploitation of the encoding by [9] to preferred sufficient reasons for the decision tree. This encoding allows for easier enumeration of preferred sufficient reasons for decision tree.

Example 3. *Let's consider the example of a banker 1 using a decision tree to decide whether to approve or reject a loan for a client. Suppose the decision tree is represented by Example 1, and the banker wants to understand why a particular instance, $\boldsymbol{x} = (1,1,1,1)$, was classified as a rejection ($T(\boldsymbol{x}) = 0$). In this case, there are multiple sufficient reasons to explain this classification. These reasons are all combinations of attributes that, if true, result in a negative classification. For $\boldsymbol{x} = (1,1,1,1)$, the sufficient reasons are: $x_1 \wedge x_2 \wedge x_4$, $x_1 \wedge x_3 \wedge x_4$, and $x_2 \wedge x_3 \wedge x_4$. However, the banker prefers an explanation without the attribute x_2 because it is a non-actionable attribute, meaning the client cannot change it. In this case, we can use a weight function for each attribute to find the best explanation. In this example, we use the weight function $w_1 = (5,1,8,4)$, which assigns higher weights to attributes considered more important for the decision. Using this weight function, the solver returns that the best explanation of maximum-weight is $x_1 \wedge x_3 \wedge x_4$, which does not include the non-actionable attribute x_2.*

5. Weight Functions and Distance Between Two Finite Subsets of Explanations

The main idea of this section is to address the variations in user preference aggregation modalities regarding preferred abductive reasons. It is acknowledged that even two experts in the same domain can have different preferences. However, in the absence of a real-world application with actual user preferences, the study focuses on exploring different weight measures, both local and global. The weight functions used in this study are based on different approaches such as Shapley values, Banzhaf values, LIME, Anchors, Explanatory, as well as Wordfreq and Feature importance. These weight functions allow quantifying the relative importance of different features

or attributes in explaining the results of the classification model. By using these weight measures, it is possible to take into account user preferences when aggregating abductive explanations, assigning different weights to features based on their perceived importance.

5.1. Weight Functions

Global Weight Measures: Global weight measures focus on the contribution of features by considering all predictions of all instances. We will present some of the global weight measures used in the literature to aggregate user preferences regarding preferred sufficient reasons.

- **Wordfreq :** Zipf's law states that the frequency f of a word in a corpus is inversely proportional to its rank r, i.e., $f \propto \frac{1}{r}$. This law is often used to model the distribution of word frequencies in a linguistic corpus. The Zipf frequency f of a word is given by: $f = \log_{10}\left(\frac{N}{r}\right)$, where N is the total number of words in the corpus and r is the rank of the word, i.e., its position in the ranking of most frequent words. [1]
- **Features importance :** The "Mean Decrease Impurity" (MDI) method is used to evaluate the importance of attributes in a classification task by measuring the average decrease in impurity (e.g., entropy or Gini index) in the decision tree when the attribute is used to divide the data into subgroups. The importance of an attribute is then evaluated by taking the average and standard deviation of this decrease in impurity over all divisions of the tree that use that attribute [17].

Local Weight Measures: Local measures focus on the contribution of features to a specific prediction, individual predicted instance. We now present some local weight measures:

- **Local Surrogate Models (LIME):** LIME allows for the explanation of individual predictions made by non-interpretable machine learning models. This technique was proposed and implemented by Ribeiro et al. in 2016 [1]. LIME focuses on constructing local surrogate models to explain individual predictions. The idea is to train an interpretable surrogate model on a new dataset composed of locally perturbed samples.
- **SHAP (SHapley Additive exPlanations):** The Shapley value is based on cooperative game theory. The goal of SHAP is to explain the prediction of an observation by calculating the contribution of each variable to that prediction. We used the method proposed by [3].
- **Anchors:** Anchors [2] is an interpretability technique that aims to find sets of rules that best summarize the behavior of the model under study. The objective is to identify the largest possible local regions where predictions are as consistent as possible.
- **Explanatory:** It involves calculating the number of models for each variable x_i given the instance x and a decision tree T using D4 [18].

Example 4. *Two other bankers have different preferences for explanations compared to the banker in Example 2. The second banker believes that if the client has not repaid a previous loan, they will never be able to repay a new loan, so they prefer an explanation with attribute x_4. These preferences are expressed with $w_2 = (1, 1, 1, 10)$. On the other hand, a third banker thinks that if the client has an annual income below $35K$ and is over 50 years old, it is preferable not to grant them a loan due to their low salary relative to their age, so they prefer an explanation with $x_2 \wedge x_4$.*

- *For $w_2 = (1, 1, 1, 10)$, the reasons $x_1 \wedge x_2 \wedge x_4$, $x_1 \wedge x_3 \wedge x_4$, and $x_2 \wedge x_3 \wedge x_4$ are preffered sufficient reasons based on the preferences of the second banker.*

[1] You can find more information at https://pypi.org/project/wordfreq/.

- *The two reasons $x_1 \wedge x_2 \wedge x_4$ and $x_2 \wedge x_3 \wedge x_4$ are two preffered sufficient reasons based based on the preferences of the third banker.*

Example 4 demonstrates that subsets of preferred reasons can be very different from each other. For instance x, the two subsets of preferred reasons based on the preferences of bankers 1 and 3 do not share any common reasons.

Monotone Transformation. We know that the operation of SAT solvers requires integer and positive weights, while the values of SHAP, LIME, etc., are not necessarily positive or integer initially. In order to satisfy this constraint for SAT solvers and still maintain the same preference order based on SHAP, LIME, etc., values, we will perform a monotonically increasing transformation on the values of different weight functions. The Explanatory method does not require a monotone transformation as the number of models for each literal is already a positive and integer value. Given a weight vector $w \in \mathbb{R}^n$, the monotone transformation is given by the following formula: $w \longleftarrow w - min_{i\in[n]}w(x_i) + 1$. Then, we multiply w by 10^k, where k is the maximum number of decimal places. This transformation allows us to convert all the weights into positive integers.

Example 5 (monotone transformation). *Let $T \in \mathtt{DT}_n$ be a decision tree and $x \in X_4$ be an instance, and let SHAP($\mathbf{x}, T$) = $(\mathbf{0.5}, -\mathbf{0.2}, \mathbf{0.3}, -\mathbf{0.1})$ be the Shapley values for the instance x given T. Then, a monotone increasing transformation gives $w(\mathbf{x}) = (\mathbf{8}, \mathbf{1}, \mathbf{6}, \mathbf{2})$.*

5.2. Distance Between Two Finite Sets of Explanations

When it comes to evaluating the impact of user preferences on preferred abductive explanations, several evaluation criteria can be considered. One of these criteria is a distance measure based on the symmetric difference between two explanations. This distance measure allows quantifying the proximity between two explanations. The symmetric difference between two explanations involves considering the literals that are present in one explanation but not in the other, that is, the literals that are specific to each explanation. By comparing the cardinality of this symmetric difference, we can assess the degree of similarity or difference between these two explanations. Additionally, we will consider the distance between two finite subsets of explanations as the minimum distance between the explanations within these two subsets.

The idea behind this distance measure is to provide an estimation of the proximity between sets of explanations, allowing us to understand how these sets come closer to or move away from each other. This can be useful for evaluating the similarities or divergences in user preferences regarding abductive explanations.

Definition 4. *The distance between two finite subsets of explanations S_1 and S_2 is defined as $S_d(S_1, S_2) = \min_{x\in S_1, y\in S_2} |D_{sr}(x, y)|$, where $|.|$ represents the counting measure, and D_{sr} is the symmetric difference between two explanations t_1 and t_2, denoted as $D_{sr}(t_1, t_2)$, given by the formula $D_{sr}(t_1, t_2) = \{l : l \in Lit(t_1) \cup Lit(t_2) \wedge l \notin Lit(t_1) \cap Lit(t_2)\} = Lit(t_1) \triangle Lit(t_2)$.*

Note that the larger the value of $S_d(S_1, S_2)$, the farther apart the two sets S_1 and S_2 are from each other. If $S_1 \cap S_2 \neq \varnothing$, then $S_d(S_1, S_2) = 0$. From a topological perspective, S_d expresses the geometric distance between two finite subsets of explanations, taking into account the topological nature of explanations, which are terms composed of literals.

Lemma 1. *The complexity of calculating the distance between two subsets of explanations, S_1 and S_2, is quadratic.*

The computational complexity of calculating the distance between two sets of explanations, S_1 and S_2, depends on the sizes of these sets. Let's assume that m_1 represents the size of S_1 and m_2 represents the size of S_2. For each element in S_1, we need to compare it with each element in S_2 to calculate the distance between them. This implies a comparison between m_1 elements of S_1 and m_2 elements of S_2, resulting in a complexity of the order of $O(m_1 \cdot m_2)$, which is quadratic when m_1 and m_2 are sufficiently large.

Example 6. *Based on Example 4, let's denote S_{w_1}, S_{w_2}, and S_{w_3} as the subsets of preferred explanations based on the preferences of bankers 1, 2, and 3, respectively. We have $S_d(S_{w_1}, S_{w_2}) = 0$ and $S_d(S_{w_2}, S_{w_3}) = 0$ because $S_{w_1} \cap S_{w_2} \neq \varnothing$ and $S_{w_3} \cap S_{w_2} \neq \varnothing$, while $S_d(S_{w_1}, S_{w_3}) = 2$.*

6. Experiments

Benchmark					Global measures				
name	$(\#B, \#I)$	$acc(\%)$	$\lvert sr(x,T)\rvert$	$\lvert sm(x,T)\rvert$	#wordf	#f_imp	$\#R_{[1,10]}$	$\#R_{[1,100]}$	$\#R_{[1,1000]}$
placement	(17,215)	86.05	9.6 (±8.87)	5.51 (±6.92)	1.84 (±1.09)	1.49 (±1.49)	2.51 (±1.23)	2.05 (±1.33)	1.67 (±1.38)
compas	(45,6172)	64.21	26.9 (±20.95)	3.26 (±2.89)	3.48 (±3.49)	2.28 (±2.28)	3.18 (±3.22)	2.18 (±1.69)	2.64 (±2.15)
diabetes	(110,768)	72.73	349 (±862)	2.4 (±2.14)	3.36 (±9.05)	3.16 (±3.16)	1.96 (±2.78)	3.3 (±4.67)	2.2 (±2.16)
indian .l	(86,583)	64.1	199 (±359)	2.06 (±1.65)	2.6 (±1.95)	2.3 (±2.3)	2.92 (±2.48)	2.32 (±2.07)	1.88 (±1.8)
banknote	(24,1372)	97.82	14.42 (±18.17)	2.76 (±2.99)	2.14 (±1.97)	2.74 (±2.74)	4.14 (±2.25)	2.16 (±1.81)	4.3 (±7.19)
anneal	(13,898)	100.0	1.28 (±0.5)	1.16 (±0.37)	3.06 (±1.35)	3.22 (±3.22)	3.18 (±1.32)	3.2 (±1.16)	3.22 (±1.13)
fetal .h	(108,2126)	93.43	10^5 ($\pm 3.10^4$)	25.5 (±25.1)	6.4 (±16.2)	2.3 (±2.34)	4.24 (±3.92)	2.92 (±3.04)	1.86 (±1.43)
divorce	(3,170)	97.06	2.29 (±0.97)	2.29 (±0.97)	1.65 (±0.95)	1.65 (±1.65)	1.85 (±0.95)	1.65 (±0.95)	1.65 (±0.95)
heart	(39,303)	70.49	15.22 (±17.36)	1.92 (±0.72)	1.74 (±1.08)	1.94 (±1.94)	1.78 (±1.6)	2.0 (±1.5)	2.06 (±1.66)
horse	(33,299)	76.67	11.98 (±14.22)	5.4 (±7.21)	2.4 (±2.56)	2.12 (±2.12)	2.42 (±3.25)	2.16 (±1.91)	2.6 (±2.32)
meta	(41,528)	88.68	61.3 (±75.11)	3.06 (±2.37)	3.82 (±5.67)	2.86 (±2.86)	3.08 (±1.95)	3.58 (±7.74)	2.7 (±2.13)
startup	(96,923)	69.19	5.10^4 ($\pm 10^6$)	44.1 (±42.3)	1.2 (±0.45)	1.6 (±1.6)	1.2 (±1.22)	1.0 (±0.6)	1.6 (±0.89)
student .p	(31,649)	90.77	21.15 (±27.88)	3.36 (±2.14)	2.62 (±2.7)	2.66 (±2.66)	2.65 (±2.32)	2.67 (±2.84)	2.51 (±2.8)
student .m	(23,395)	86.08	7.53 (±7.1)	1.78 (±1.74)	2.78 (±2.76)	2.49 (±2.49)	2.15 (±2.28)	1.99 (±1.35)	2.18 (±1.66)
derm	(4,366)	98.65	2.23 (±1.23)	2.18 (±1.26)	2.54 (±0.98)	2.54 (±2.54)	1.85 (±0.98)	1.46 (±0.5)	1.36 (±0.48)
balance	(17,625)	87.2	27.2 (±32.99)	2.8 (±2.2)	1.9 (±1.1)	3.0 (±3.0)	2.3 (±2.38)	2.9 (±4.01)	2.2 (±1.14)
hepatitis	(15,142)	82.76	6.52 (±3.93)	2.45 (±1.43)	1.9 (±0.86)	2.28 (±2.28)	1.83 (±0.73)	2.07 (±1.1)	1.24 (±0.64)

Figure 2: Statistics on the computation of preferred sufficient reasons using a global measures for instances from various datasets.

Experimental setup. We considered 18 well-known binary classification datasets available on Kaggle, OpenML, and UCI. No data preprocessing was performed for numerical attributes, and the attributes were binarized in-line by the decision tree learning algorithm used. For each benchmark b, we evaluated the classification performance using standard evaluation metrics. We used the CART algorithm and its implementation in Scikit-Learn to learn decision trees, with default parameter settings. For each benchmark b and a subset of up to 250 randomly selected instances $\boldsymbol{x}$ from the test set, unless the dataset contains fewer than 250 instances, in which case the entire dataset was used. We computed the number of sufficient reasons using the encoding proposed by [9] and the number of minimum-size sufficient reasons using the Partial MaxSAT solver (with a 60-second timeout per instance). Finally, we computed the number of preferred sufficient reasons using the encoding detailled in the section 4 and the **WEIGHTED PARTIAL MAXSAT** solver from OpenWBO [19].

Regarding the weight functions, for each tree T_b, we used the exact method proposed by [20] to compute the **SHAP** score as well as the scores for **LIME** [1] and **Anchors** [2]. We also used feature importance with Scikit-Learn [17], the number of models "**Explanatory**" with [18], and the **Zipf frequency** of each feature viewed as a word in the wordfreq library. Two weight functions (random local and global) based on random weight sampling were added to clarify the

nature of preferred explanations for different weight functions. We report the classical statistics, the average number and variance of sufficient reasons and minimal sufficient reasons, and the preferred sufficient reasons for each weight function method. Finally, for the "**placement**" and "**compas**" datasets, we report the distance between different preferred subsets for the different weight functions.

Benchmark					Global measures				
name	$(\#B, \#I)$	$acc(\%)$	$\lvert sr(x,T)\rvert$	$\lvert sm(x,T)\rvert$	#wordf	#f_imp	$\#R_{[1,10]}$	$\#R_{[1,100]}$	$\#R_{[1,1000]}$
placement	(17,215)	86.05	9.6 (±8.87)	5.51 (±6.92)	1.84 (±1.09)	1.49 (±1.49)	2.51 (±1.23)	2.05 (±1.33)	1.67 (±1.38)
compas	(45,6172)	64.21	26.9 (±20.95)	3.26 (±2.89)	3.48 (±3.49)	2.28 (±2.28)	3.18 (±3.22)	2.18 (±1.69)	2.64 (±2.15)
diabetes	(110,768)	72.73	349 (±862)	2.4 (±2.14)	3.36 (±9.05)	3.16 (±3.16)	1.96 (±2.78)	3.3 (±4.67)	2.2 (±2.16)
indian .l	(86,583)	64.1	199 (±359)	2.06 (±1.65)	2.6 (±1.95)	2.3 (±2.3)	2.92 (±2.48)	2.32 (±2.07)	1.88 (±1.8)
banknote	(24,1372)	97.82	14.42 (±18.17)	2.76 (±2.99)	2.14 (±1.97)	2.74 (±2.74)	4.14 (±2.25)	2.16 (±1.81)	4.3 (±7.19)
anneal	(13,898)	100.0	1.28 (±0.5)	1.16 (±0.37)	3.06 (±1.35)	3.22 (±3.22)	3.18 (±1.32)	3.2 (±1.16)	3.22 (±1.13)
fetal .h	(108,2126)	93.43	10^5 ($\pm 3.10^4$)	25.5 (±25.1)	6.4 (±16.2)	2.3 (±2.34)	4.24 (±3.92)	2.92 (±3.04)	1.86 (±1.43)
divorce	(3,170)	97.06	2.29 (±0.97)	2.29 (±0.97)	1.65 (±0.95)	1.65 (±1.65)	1.85 (±0.95)	1.65 (±0.95)	1.65 (±0.95)
heart	(39,303)	70.49	15.22 (±17.36)	1.92 (±0.72)	1.74 (±1.08)	1.94 (±1.94)	1.78 (±1.6)	2.0 (±1.5)	2.06 (±1.66)
horse	(33,299)	76.67	11.98 (±14.22)	5.4 (±7.21)	2.4 (±2.56)	2.12 (±2.12)	2.42 (±3.25)	2.16 (±1.91)	2.6 (±2.32)
meta	(41,528)	88.68	61.3 (±75.11)	3.06 (±2.37)	3.82 (±5.67)	2.86 (±2.86)	3.08 (±1.95)	3.58 (±7.74)	2.7 (±2.13)
startup	(96,923)	69.19	5.10^4 ($\pm 10^6$)	44.1 (±42.3)	1.2 (±0.45)	1.6 (±1.6)	1.2 (±1.22)	1.0 (±0.6)	1.6 (±0.89)
student .p	(31,649)	90.77	21.15 (±27.88)	3.36 (±2.14)	2.62 (±2.7)	2.66 (±2.66)	2.65 (±2.32)	2.67 (±2.84)	2.51 (±2.8)
student .m	(23,395)	86.08	7.53 (±7.1)	1.78 (±1.74)	2.78 (±2.76)	2.49 (±2.49)	2.15 (±2.28)	1.99 (±1.35)	2.18 (±1.66)
derm	(4,366)	98.65	2.23 (±1.23)	2.18 (±1.26)	2.54 (±0.98)	2.54 (±2.54)	1.85 (±0.98)	1.46 (±0.5)	1.36 (±0.48)
balance	(17,625)	87.2	27.2 (±32.99)	2.8 (±2.2)	1.9 (±1.1)	3.0 (±3.0)	2.3 (±2.38)	2.9 (±4.01)	2.2 (±1.14)
hepatitis	(15,142)	82.76	6.52 (±3.93)	2.45 (±1.43)	1.9 (±0.86)	2.28 (±2.28)	1.83 (±0.73)	2.07 (±1.1)	1.24 (±0.64)

Figure 3: Statistics on the computation of preferred sufficient reasons using a local measures for instances from various datasets

6.1. Experimental results

Tables 2 and 3 present an excerpt of the results. The tables present results on datasets, decision trees, and global weight measures, based on 18 datasets. For each benchmark, the table provides the dataset name (name), the accuracy of the decision trees $(acc(\%))$, the number of binary variables $(\#B)$, and the number of instances $(\#I)$. The columns $|sr(x,T)|$ and $|sm(x,T)|$ respectively indicate the mean and standard deviation (std) of the number of sufficient reasons and the number of preferred sufficient reasons. Then, for each benchmark b, the columns #wordf, #f_imp, $(R_{[1,10]}, R_{[1,100]}, R_{[1,1000]})$ correspondingly represent the number of preferred sufficient reasons for wordfreq, feature importance, and global random sampling over the intervals [1,10], [1,100], and [1,1000]. The columns of Table 3 represent the mean and standard deviation (std) of the number of preferred sufficient reasons for the local weight measures in the following order: Lime, Shapely, Anchors, Explanatory, and local random sampling over the intervals [1,10], [1,100], and [1,1000]. We clarify that the concept of "random sampling local" consists of selecting integer weights for each instance, while respecting a specified interval. Let's consider the illustrative example: suppose we have a dataset with instances of size $n = 5$, meaning that there are five elements in each instance. The specified interval is [1, 10], indicating that the chosen weights must be integer values ranging from 1 to 10. For each individual instance, we perform a random draw to determine the corresponding weights. In our example, the weight vector $w = (9, 4, 7, 5)$ is generated from this random draw. Each weight in the vector is an integer chosen randomly within the interval [1, 10].

First. We would like to emphasize that computing preferred reasons given a decision tree and instance is feasible in practice. In fact, for many datasets and instances, the computation of all preferred reasons has been completed in less than 20 seconds, regardless of the type of weight

/	Lime	Shap	Anchor	Exp	"R_[1,10]"	"R_[1,100]"	"R_[1,1000]"
Lime	0.0	0.16	0.2	0.28	0.2	0.28	0.38
Shap	.	0.0	0.32	0.4	0.32	0.4	0.5
Anchor	.	.	0.0	0.32	0.22	0.3	0.4
Exp	.	.	.	0.0	0.32	0.4	0.5
"R_[1,10]"	.	.	.	.	0.0	0.3	0.4
"R_[1,100]"	.	.	.	.	.	0.0	0.36
"R_[1,1000]"	.	.	.	.	.	.	0.0

Figure 4: Statistics on the symmetric distance between the subsets of preferred sufficient reasons using local measures for the compas dataset.

/	Lime	Shap	Anchor	Exp	"R_[1,10]"	"R_[1,100]"	"R_[1,1000]"
Lime	0.0	0.17	0.23	0.01	0.23	0.17	0.2
Shap	.	0.0	0.23	0.0	0.23	0.2	0.2
Anchor	.	.	0.0	0.23	0.23	0.23	0.23
Exp	.	.	.	0.0	0.23	0.17	0.2
"R_[1,10]"	.	.	.	.	0.0	0.23	0.23
"R_[1,100]"	.	.	.	.	.	0.0	0.2
"R_[1,1000]"	.	.	.	.	.	.	0.0

Figure 5: Statistics on the symmetric distance between the subsets of preferred sufficient reasons using local measures for the placement dataset.

function used. It is evident that the use of different weight function types has a significant impact on the number of reasons, making it easier to compute all preferred reasons by reducing their quantity compared to sufficient reasons and minimum-size reasons.

Furthermore, it is important to note that for each dataset b, each instance in the benchmark of b, and each type of weight function, enumerating the preferred sufficient reasons has been feasible. Leveraging user preferences offers a significant advantage by substantially reducing the number of generated explanations. By focusing solely on the explanations preferred by the user, information overload is avoided, and attention is directed towards the most relevant and useful explanations.

Second. Tables 4 and 5 present a matrix that visualizes the average distances between different subsets of explanations. These subsets of explanations are obtained using various methods of local and global weight assignment. The values in the matrices correspond to the distances between pairs of subsets, where the coordinates (x, y) represent the weight assignment methods used. When examining the diagonal entries of the matrix, we observe that the distances are zero. This is because a subset is identical to itself, so the distance between a subset and itself is always 0. Additionally, it is important to note that the matrices are symmetric. This is because the distance used is symmetric, which is typically the case for all distances.

By observing the distances between the different subsets of explanations, we notice that they are generally less than 1. This indicates that the explanations are relatively close to each other in terms of distance. Topologically, this suggests that the set of sufficient reasons forms a compact structure, where the explanations are closely grouped and interconnected. This observation represents an initial step in studying the diversity of formal explanations. It indicates that the

different methods of local and global weight assignment used to generate the explanations do not result in explanations that are very distant from each other. This raises questions about the variety and extent of possible explanations, as well as how local weight assignment methods can influence the diversity of the obtained explanations.

7. Conclusion

To summarize the contributions highlighted in this article, we first proposed a CNF-encoding approach to compute preferred sufficient reasons for decision trees. This approach involves representing the reasons in a logical form that facilitates their calculation. Additionally, we introduced the concept of distance between preferred explanations and examined the impact of weight functions on preferred abductive explanations. Namely, we investigated how different methods of assigning weights affect the proximity of preferred explanations to each other. Our focus was on the quantity and diversity of these explanations. We found that a classified instance, whether positive or negative, can have an exponential number of reasons, including an exponential number of minimum-sized reasons or preferred reasons. This means that there can be numerous possible explanations for a single classified instance. However, despite this potential diversity, the number of preferred reasons is significantly smaller than the number of sufficient reasons, regardless of the weight function used. Generally, there is a restricted selection of preferred explanations that are considered the most relevant or useful. Furthermore, we observed that the distances between different sets of explanations are generally not large. This indicates that abductive explanations for decision trees tend to be close to each other in terms of similarity or proximity. In other words, the explanations often share similar features or partially overlap. These findings suggest that despite the potential diversity of explanations, there are commonalities and trends among preferred explanations for decision trees. This can be useful in understanding how decisions are made by these models and in providing comprehensible explanations to users.

Studying the impact of weight functions on preferred abductive explanations for decision trees is just the first step in our research on the diversity of abductive explanations. We intend to apply a similar approach to other models, particularly random forests. Concurrently, we are developing a SAT encoding to compute the SAT Distance between preferred sets of sufficient reasons. The aim of this endeavor is to provide users with a framework for selecting preferred explanations that align with their personal preferences and are closer to the model's output. In other words, through this SAT encoding, users will be able to measure the proximity between different sets of explanations and identify those that are most relevant and consistent with their expectations. This will enhance their understanding of the model's results and enable the provision of explanations that are better suited to the users needs.

References

[1] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, in: Proc. of SIGKDD'16, 2016, pp. 1135–1144.

[2] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proc. of AAAI'18, 2018, pp. 1527–1535.

[3] S. Lundberg, S.-I. Lee, A unified approach to interpreting m(ijcaiodel predictions, in: Proc. of NIPS'17, 2017, pp. 4765–4774.

[4] C. Molnar, Interpretable Machine Learning, Leanpub, 2020.

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

[6] Y. Izza, A. Ignatiev, J. Marques-Silva, On explaining decision trees, CoRR abs/2010.11034 (2020).

[7] A. Shih, A. Choi, A. Darwiche, A symbolic approach to explaining bayesian network classifiers, in: Proc. of IJCAI'18, 2018, pp. 5103–5111.

[8] A. Darwiche, A. Hirth, On the reasons behind decisions, in: Proc. of ECAI'20, 2020.

[9] S. Jabbour, J. Marques-Silva, L. Sais, Y. Salhi, Enumerating prime implicants of propositional formulae in conjunctive normal form, in: Logics in Artificial Intelligence, 2014.

[10] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J.-M. Lagniez, P. Marquis, On the explanatory power of boolean decision trees, Data & Knowledge Engineering 142 (2022) 102088. URL: https://www.sciencedirect.com/science/article/pii/S0169023X22000799. doi:https://doi.org/10.1016/j.datak.2022.102088.

[11] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J.-M. Lagniez, P. Marquis, Trading complexity for sparsity in random forest explanations, in: Proc. of AAAI'22, 2022.

[12] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J.-M. Lagniez, P. Marquis, Sur le pouvoir explicatif des arbres de décision, EGC'2022 38 (2022).

[13] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J. Lagniez, P. Marquis, On preferred abductive explanations for decision trees and random forests, in: Proc. of IJCAI'22, 2022.

[14] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, 2017. arXiv:1702.08608.

[15] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J.-M. Lagniez, P. Marquis, Les raisons majoritaires: des explications abductives pour les forêts aléatoires, EGC'2022 38 (2022).

[16] G. Audemard, S. Bellart, L. Bounia, F. Koriche, J. Lagniez, P. Marquis, On the computational intelligibility of boolean classifiers, in: Proc. of KR'21, 2021, pp. 74–86.

[17] 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, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[18] J.-M. Lagniez, P. Marquis, An Improved Decision-DNNF Compiler, in: Proc. of IJCAI'17, 2017, pp. 667–673.

[19] R. Martins, V. M. Manquinho, I. Lynce, Open-wbo: A modular maxsat solver,, in: International Conference on Theory and Applications of Satisfiability Testing, 2014.

[20] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, Explainable ai for trees: From local explanations to global understanding, arXiv preprint arXiv:1905.04610 (2019).