Leveraging Argumentation for Generating Robust Sample-based Explanations

Leila Amgoud[1] , **Philippe Muller**[2] and **Henri Trenquier**[3]
[1]CNRS – IRIT
[2]Toulouse University – IRIT
[3]Toulouse University – ANITI
{leila.amgoud, philippe.muller}@irit.fr, henri.trenquier@univ-tlse3.fr

Abstract

Explaining predictions made by inductive classifiers has become crucial with the rise of complex models acting more and more as black-boxes. *Abductive explanations* are one of the most popular types of explanations that are provided for the purpose. They highlight feature-values that are sufficient for making predictions. In the literature, they are generated by exploring the whole feature space, which is unreasonable in practice. This paper solves the problem by introducing explanation functions that generate abductive explanations from a sample of instances. It shows that such functions should be defined with great care since they cannot satisfy two desirable properties at the same time, namely existence of explanations for every individual decision (success) and correctness of explanations (coherence). The paper provides a parameterized family of argumentation-based explanation functions, each of which satisfies one of the two properties. It studies their formal properties and their experimental behaviour on different datasets.

1 Introduction

Recent advances in many AI fields rely on inductive models, depending on parameters that are adjusted based on a set of training instances. Such models tend to be large for practical tasks, in the sense of having a lot of parameters, and may allow for non-linear interactions between input features. Consequently, they are perceived as *black-boxes* whose behaviour is difficult to grasp both from their designers and users' point of view. This opacity has sparked a new sub-field of AI, *explainable AI* (XAI), whose approaches provide ways to explain what black-box models do and why they do it (see [Burkart and Huber, 2021] for a recent survey on XAI).

One of the most studied types of explanation is the so-called *abductive explanations*, which highlight feature-values that are sufficient for making a given prediction. For example, a client was refused a loan *because he is unemployed*. Such explanations are generally generated from the whole feature space (eg., [Darwiche and Hirth, 2020; Ignatiev *et al.*, 2019b; Audemard *et al.*, 2022; Amgoud, 2021a]). While the approach is reasonable when models are interpretable, like Decision Trees or Random Forests, it is not tractable in case of black-boxes (see [Cooper and Marques-Silva, 2021]) as it requires an exhaustive exploration of the feature space.

As a solution, the two prominent explanation functions Anchors [Ribeiro *et al.*, 2018] and LIME [Ribeiro *et al.*, 2016] and the argument-based function [Amgoud, 2021b] generate abductive explanations from a sample (i.e., subset) of instances, avoiding thus exploring the whole feature space. However, it has been shown in [Amgoud, 2021b; Narodytska *et al.*, 2019] that the explanations of Anchors/LIME may be globally inconsistent and thus incorrect. The third function ensures correct explanations but does not guarantee the existence of explanations for every instance. Furthermore, it is very cautious as it simply removes all conflicting explanations that may be generated from the considered sample.

This paper investigates explanation functions that generate abductive explanations from a subset of feature space while satisfying desirable properties. Its contributions are fourfold:

The first consists of proving an impossibility result, which states that a function that generates abductive explanations from a subset of instances cannot guarantee both existence of explanations (success) and their correctness (coherence). This result sheds light on the reason behind violation of success by the argument-based function from [Amgoud, 2021b].

The second contribution consists of a parameterized family of argumentation-based explanation functions, each of which satisfies one of the two incompatible properties. The approach starts by generating arguments in favour of classes, identifies attacks among them, uses stable semantics [Dung, 1995] for generating sets of arguments that can be jointly accepted, identifies *accepted arguments*, and uses the latter for defining the novel types of abductive explanations. Accepted arguments are defined in our approach using two parameters: *selection function* and *inference rule*. The former selects a subset of stable extensions and the latter selects (accepted) arguments from the chosen extensions. We define various instantiations of the two parameters, capturing different *criteria* for solving conflicts between arguments.

The third contribution is a formal analysis and a comprehensive comparison of the new functions. We show that the family encompasses the argument-based function, however the new functions that ensure correctness of explanations perform better as they explain more instances and more classes.

The fourth contribution is an experimental analysis of the functions on various datasets. The results confirm that abductive explanations that are generated from datasets (as done by Anchors) are generally incorrect. They show also that the new functions which guarantee correctness perform well as they explain quite an important proportion of instances.

The paper is organized as follows: Section 2 gives some background, Section 3 defines plausible explanations, Section 4 presents two desirable properties and shows their incompatibility, Section 5 defines the novel family of argumentation-based functions, and studies their properties, Section 6 presents the first experimental results, Section 7 is devoted to related work, and the last section concludes.

2 Preliminaries

Throughout the paper, we consider a classification theory as a tuple made of a finite set of *features*, a function which returns the *domain* of every feature and a finite set of *classes*.

Definition 1 (Theory)**.** *A* theory *is a tuple* $\mathtt{T} = \langle \mathtt{F}, \mathtt{D}, \mathtt{C} \rangle$ *s.t.*

- $\mathtt{F}$ *is a finite set of features,*
- $\mathtt{D}$ *is a function on* $\mathtt{F}$ *such that, for every* $f \in \mathtt{F}$, $\mathtt{D}(f)$ *is countable (discrete domains),*
- $\mathtt{C}$ *is a finite set of possible distinct classes with* $|\mathtt{C}| > 1$.

We introduce next the useful notion of literal.

Definition 2 (Literal)**.** *Let* $\mathtt{T} = \langle \mathtt{F}, \mathtt{D}, \mathtt{C} \rangle$ *be a theory. A* literal *is a pair* (f, v) *where* $f \in \mathtt{F}$ *and* $v \in \mathtt{D}(f)$. *Let* $\mathtt{Lit}(\mathtt{T})$ *denote the set of all possible literals of* $\mathtt{T}$.

A set of literals is consistent if it does not contain two literals having the same attribute but distinct values.

Definition 3 (Consistency)**.** *A set* $L \subseteq \mathtt{Lit}(\mathtt{T})$ *is* consistent *iff* $\nexists (f, v), (f', v') \in L$ *such that* $f = f'$ *and* $v \neq v'$. *Otherwise,* L *is said to be* inconsistent.

We call *instance* any assignment of values to all features.

Definition 4 (Instance)**.** *Let* $\mathtt{T} = \langle \mathtt{F}, \mathtt{D}, \mathtt{C} \rangle$ *be a theory. An* instance *is a subset* I *of literals such that every attribute* $f \in \mathtt{F}$ *appears exactly once in* I. *Let* $\mathbb{I}_{\mathtt{T}}$ *denote the set of all instances of* $\mathtt{T}$, *called* feature space.

A classification model, or classifier, is a surjective function mapping every instance into a single prediction.

Definition 5 (Classifier)**.** *Let* $\mathtt{T} = \langle \mathtt{F}, \mathtt{D}, \mathtt{C} \rangle$ *be a theory. A* classifier *on* $\mathtt{T}$ *is a surjective function* $\mathtt{R}$ *from* $\mathbb{I}_{\mathtt{T}}$ *to* $\mathtt{C}$.

An explanation function answers questions of the form: why does classifier $\mathtt{R}$ assign class c to instance x? One of the most studied types of reasons in the AI literature for a long time is abductive explanations (eg., [Dimopoulos *et al.*, 1997; Kakas and Riguzzi, 2000]). More recently, they have been used for interpreting classifiers (eg., [Shih *et al.*, 2018; Ignatiev *et al.*, 2019a; Darwiche and Hirth, 2020]). An abductive explanation is defined as a subset-minimal set of literals that is sufficient for predicting the class of an instance.

Definition 6 (Abductive Explainer)**.** *Let* $\mathtt{R}$ *be a classifier and* $\mathtt{T}$ *a theory. An* abductive explainer *is a function* $\mathtt{g}_a$ *mapping every* $I \in \mathbb{I}_{\mathtt{T}}$ *into the set of any* L *verifying the following:*

a) $L \subseteq I$,

b) $\forall I' \in \mathbb{I}_{\mathtt{T}} \setminus \{I\}$ *such that* $L \subseteq I'$, $\mathtt{R}(I') = \mathtt{R}(I)$,

c) $\nexists L' \subset L$ *such that* L' *satisfies the above conditions.*

The set of literals L *is called* abductive explanation.

Every instance may have one or several abductive explanations as shown in the following example.

Example 1. *Consider a theory made of two binary features* f_1, f_2 *and three classes* c_1, c_2, c_3. *The table below summarizes the predictions made by a classifier* $\mathtt{R}$.

$\mathbb{I}(\mathtt{T})$	f_1	f_2	$\mathtt{R}(I_i)$
I_1	*0*	*0*	c_1
I_2	*0*	*1*	c_2
I_3	*1*	*0*	c_3
I_4	*1*	*1*	c_3

The abductive explanations of I_1, I_2, I_3, I_4 *are given below.*

- $\mathtt{g}_a(I_1) = \{L_1\}$ $\qquad L_1 = \{(f_1, 0), (f_2, 0)\}$
- $\mathtt{g}_a(I_2) = \{L_2\}$ $\qquad L_2 = \{(f_1, 0), (f_2, 1)\}$
- $\mathtt{g}_a(I_3) = \mathtt{g}_a(I_4) = \{L_3\}$ $\qquad L_3 = \{(f_1, 1)\}$

The condition b) in the above definition states that generating an abductive explanation for the prediction of an instance requires testing a set L of literals on the whole feature space, especially when the classifier $\mathtt{R}$ is a black-box. This is not reasonable due to the huge size of the feature space in practice, and the complexity of querying black-box classifiers like deep neural networks. Indeed, it has been shown in [Cooper and Marques-Silva, 2021] that the complexity of finding one abductive explanation in case of black-box classifiers is co-NP-complete. In what follows we propose an alternative solution, which consists of testing L only on a subset of instances.

3 Plausible Abductive Explanations

In the remaining of the paper, we assume fixed but arbitrary theory $\mathtt{T} = \langle \mathtt{F}, \mathtt{D}, \mathtt{C} \rangle$ and black-box classifier $\mathtt{R}$. We also consider a sample $\mathcal{Y} \subseteq \mathbb{I}_{\mathtt{T}}$ of the feature space. This set may represent a dataset on which $\mathtt{R}$ is trained or may be generated in specific ways. Whatever its source, it should satisfy a property stating that every class in $\mathtt{C}$ should be represented in $\mathcal{Y}$ (i.e., $\forall c \in \mathtt{C}$, $\exists I \in \mathcal{Y}$ such that $\mathtt{R}(I) = c$). This condition ensures a quite well-balanced sample. Furthermore, as we will show later, explanations generated from a sample are approximations of those generated from the feature space. Hence, the condition increases the quality of approximations. We call the latter *plausible abductive explanations*.

Definition 7 (Plausible Explainer)**.** *Let* $\mathtt{R}$ *be a classifier,* $\mathtt{T}$ *a theory,* $\mathcal{Y} \subseteq \mathbb{I}_{\mathtt{T}}$. *A* plausible explainer *is a function* $\mathtt{g}_p$ *mapping every* $I \in \mathcal{Y}$ *into the set of any* L *verifying the following:*

a) $L \subseteq I$,

b) $\forall I' \in \mathcal{Y} \setminus \{I\}$ *such that* $L \subseteq I'$, $\mathtt{R}(I') = \mathtt{R}(I)$,

c) $\nexists L' \subset L$ *such that* L' *satisfies the above conditions.*

The set L *is called* plausible abductive explanation.

Let us illustrate the definition on an example.

Example 2. *Let us consider the theory made of four binary features and three classes. Assume a classifier* $\mathtt{R}$ *which provides the predictions below for the seven instances in* $\mathcal{Y}$.

$\mathcal{Y}$	f_0	f_1	f_2	f_3	$\mathtt{R}(I_i)$
I_1	*0*	*0*	*1*	*0*	c_1
I_2	*0*	*0*	*1*	*1*	c_1
I_3	*0*	*1*	*0*	*0*	c_0
I_4	*0*	*1*	*0*	*1*	c_2
I_5	*0*	*1*	*1*	*1*	c_1
I_6	*1*	*1*	*0*	*1*	c_2
I_7	*1*	*1*	*1*	*0*	c_2

The function $\mathtt{g}_p$ *returns the following explanations.*

- $\mathtt{g}_p(I_1) = \{L_5, L_7\}$ $L_1 = \{(f_2, 0), (f_3, 0)\}$
- $\mathtt{g}_p(I_2) = \{L_2, L_5, L_7\}$ $L_2 = \{(f_2, 1), (f_3, 1)\}$
- $\mathtt{g}_p(I_3) = \{L_1, L_6\}$ $L_3 = \{(f_0, 1)\}$
- $\mathtt{g}_p(I_4) = \{L_4\}$ $L_4 = \{(f_2, 0), (f_3, 1)\}$
- $\mathtt{g}_p(I_5) = \{L_2, L_7\}$ $L_5 = \{(f_1, 0)\}$
- $\mathtt{g}_p(I_6) = \{L_3, L_4\}$ $L_6 = \{(f_0, 0), (f_1, 1), (f_3, 0)\}$
- $\mathtt{g}_p(I_7) = \{L_3, L_8\}$ $L_7 = \{(f_0, 0), (f_2, 1)\}$
- $L_8 = \{(f_1, 1), (f_2, 1), (f_3, 0)\}$

We show[1] that every abductive explanation of an instance is a superset of a plausible explanation of the same instance. This shows that plausible explanations are approximations of and shorter than abductive ones.

Proposition 1. *Let* $\mathtt{T}$ *be a theory and* $\mathcal{Y} \subseteq \mathbb{I}_{\mathtt{T}}$. *For every* $I \in \mathcal{Y}$, *if* $L \in \mathtt{g}_a(I)$, *then* $\exists L' \subseteq L$ *such that* $L' \in \mathtt{g}_p(I)$.

The following example shows that the converse does not hold, i.e., a plausible explanation may not be the subset of any abductive explanation provided by the function $\mathtt{g}_a$.

Example 2 (Cont.) Assume that the prediction of the instance $I_8 \in \mathbb{I}_{\mathtt{T}}$ below is $\mathtt{R}(I_8) = c_1$.

	f_0	f_1	f_2	f_3	$\mathtt{R}(I_8)$
I_8	1	1	0	0	c_1

Note that $L_1 \in \mathtt{g}_p(I_3)$ while L_1 cannot be a subset of an abductive explanation (i.e., $L_0 \notin \mathtt{g}_a(I_3)$).

It is worth mentioning that generating one plausible explanation can be achieved in polynomial time. It depends simply on the number of instances in the sample and the number of features in the theory. While this gain is important, we show next that the plausible explainer suffers from a tricky issue.

4 Coherence vs Existence of Explanations

In [Amgoud and Ben-Naim, 2022], the authors introduced a set of principles for explanation functions that interpret the global behaviour of a classifier, i.e., those that explain classes instead of instances. Every principle is seen as a desirable property that should be satisfied. In what follows we adapt two of them to functions explaining instances from samples.

Definition 8. *Let* $\mathtt{R}$ *be a classifier,* $\mathtt{T}$ *a theory and* $\mathcal{Y} \subseteq \mathbb{I}_{\mathtt{T}}$. *A* refined plausible explainer *is a function* $\mathtt{g}$ *mapping every* $I \in \mathcal{Y}$ *into* $\mathtt{g}(I) \subseteq \mathtt{g}_p(I)$.

The first principle, called *success*, states that any refined plausible explainer should return at least one explanation to every instance. It ensures feedback for end-users.

Principle 1. *(Success) A refined plausible explainer* $\mathtt{g}$ *satisfies* success *iff for any classifier* $\mathtt{R}$, *any theory* $\mathtt{T}$, *any* $\mathcal{Y} \subseteq \mathbb{I}_{\mathtt{T}}$, *and any* $I \in \mathcal{Y}$, *we have that* $\mathtt{g}(I) \neq \emptyset$.

The second principle, called *coherence*, states that the explanations of instances labelled with different classes should be inconsistent. This property prevents the following three undesirable situations: Assume two instances $I, I' \in \mathcal{Y}$ such that $\mathtt{R}(I) \neq \mathtt{R}(I')$. Assume also that L is an explanation for I and L' is an explanation for I'. We may have the following:

i) $L = L'$,

ii) $L \subset L'$,

iii) $L \not\subseteq L'$ and $L \cup L'$ is consistent.

It is clearly not reasonable to predict different classes on the basis of the same set of information (i), ii)). For the third case, assume L and L' stand respectively for: Age ≤ 45, salary $\leq 50K$ and $\mathtt{R}(I)$ and $\mathtt{R}(I')$ stand for accepting and rejecting a loan respectively. The two explanations are incompatible since they both match a profile of a customer whose age is 30 and salary is 40K. The first rule states that this customer should have the loan while the second predicts rejection.

Principle 2. *(Coherence) A refined plausible explainer* $\mathtt{g}$ *satisfies* coherence *iff for any classifier* $\mathtt{R}$, *any theory* $\mathtt{T}$, *any* $I, I' \in \mathcal{Y}$, *if* $\mathtt{R}(I) \neq \mathtt{R}(I')$, *then* $\forall L \in \mathtt{g}(I)$, $\forall L' \in \mathtt{g}(I')$, $L \cup L'$ *is inconsistent.*

It is well-known in the literature that the function $\mathtt{g}_a$ provides at least one explanation for each instance in the theory's feature space. From Proposition 1, it follows that the same holds for the plausible explainer $\mathtt{g}_p$, thus $\mathtt{g}_p$ satisfies success.

Proposition 2. *For any theory* $\mathtt{T}$, *any* $\mathcal{Y} \subseteq \mathbb{I}_{\mathtt{T}}$, *any classifier* $\mathtt{R}$, *and any* $I \in \mathcal{Y}$, $\mathtt{g}_p(I) \neq \emptyset$.

The situation is different for the second principle. Indeed, the following example shows that the plausible explainer $\mathtt{g}_p$ violates coherence, and may provide **erroneous** explanations.

Example 2 (Cont.) Consider the two instances I_2 and I_3. Note that $\mathtt{R}(I_2) \neq \mathtt{R}(I_3)$ while $L_5 \in \mathtt{g}_p(I_2)$, $L_1 \in \mathtt{g}_p(I_3)$ and $L_1 \cup L_5$ is consistent. Consequently, there exists $I' \in \mathbb{I}_{\mathtt{T}}$ such that $L_1 \cup L_5 \subseteq I'$. Since I' has a single class, then at least one of the two explanations (L_1, L_5) is incorrect.

In what follows, we show that the two principles (success, coherence) are *incompatible* when explanations are generated from a dataset or more generally from a subset of instances. In other words, there is no (refined) plausible explainer that can satisfy the two principles at the same time for every classifier, every theory, and every subset of the feature space.

Theorem 1. *There is no refined plausible explainer that satisfies both coherence and success.*

To sum up, the previous result shows that generating abductive explanations from a subset of feature space is a tricky issue. A refined plausible explainer can either, like $\mathtt{g}_p$ always guarantee explanations for every instance but they may be wrong, or provide correct explanations for only a subset of instances. The following section defines in a **unified setting** various functions for each of the two policies.

[1] N.B. All proofs are listed in the technical annex because of space constraints.

5 Parameterized Family of Explainers

Throughout this section we consider an arbitrary but fixed subset $\mathcal{Y} \subseteq \mathbb{I}_{\mathtt{T}}$ of instances of theory $\mathtt{T} = \langle \mathtt{F}, \mathtt{D}, \mathtt{C} \rangle$. We define a novel parameterized family of refined explanation functions. The family is based on argumentation theory (see [Rahwan and Simari, 2009] on more on argumentation) and follows thus the following steps: it starts by generating arguments from $\mathcal{Y}$, identifies attacks among them, uses a semantics for generating sets of arguments that can be jointly accepted, identifies accepted arguments, and uses the latter for defining novel types of abductive explanations.

In our approach, arguments support classes, in the sense that they provide minimal sets of literals that determine a class. They are thus independent from instances. An advantage of not considering instances is to reduce the number of arguments that can be built from $\mathcal{Y}$. Furthermore, explanations of an instance are explanations of its predicted class.

Definition 9 (Argument)**.** *An* argument *built from $\mathcal{Y}$ is a pair $\langle L, c \rangle$ such that:*

- $L \subseteq \mathtt{Lit}(\mathtt{T})$ *and* $c \in \mathtt{C}$*,*
- $\exists I \in \mathcal{Y}$ *such that* $L \subseteq I$*,*
- $\forall I \in \mathcal{Y}$ *such that* $L \subseteq I$*,* $\mathtt{R}(I) = c$*,*
- $\nexists L' \subset L$ *that verifies the above conditions.*

L and c are called respectively support *and* conclusion *of the argument.* $\mathtt{Arg}(\mathcal{Y})$ *denotes the set of arguments built from $\mathcal{Y}$.*

The second condition of the above definition ensures that arguments are extracted from instances of the set $\mathcal{Y}$. It discards any fallacious argument whose support is not included in any instance of $\mathcal{Y}$ and thus satisfies the third condition in a vacuous way. The third condition states that the support L is correlated to the conclusion c.

Example 2 (Cont.) Eight arguments are generated from $\mathcal{Y}$:

- $a_1 = \langle L_1, c_0 \rangle \quad a_2 = \langle L_6, c_0 \rangle$
- $a_3 = \langle L_2, c_1 \rangle \quad a_4 = \langle L_5, c_1 \rangle \quad a_5 = \langle L_7, c_1 \rangle$
- $a_6 = \langle L_3, c_2 \rangle \quad a_7 = \langle L_4, c_2 \rangle \quad a_8 = \langle L_8, c_2 \rangle$

Notice that the support of every argument is a plausible abductive explanation of one or more instances in $\mathcal{Y}$. Before presenting the result, let us first introduce two notations.

Notations: Let $\mathcal{E} \subseteq \mathtt{Arg}(\mathcal{Y})$. We denote by $\mathtt{cov_i}(\mathcal{E})$ the set of instances covered by $\mathcal{E}$, ie., $\mathtt{cov_i}(\mathcal{E}) = \{I \in \mathcal{Y} \mid \exists \langle L, c \rangle \in \mathcal{E}$ and $L \subseteq I\}$, and by $\mathtt{cov_c}(\mathcal{E})$ the set of classes covered by $\mathcal{E}$, ie., $\mathtt{cov_c}(\mathcal{E}) = \{c \in \mathtt{C} \mid \exists \langle L, c \rangle \in \mathcal{E}\}$.

Proposition 3. *The following properties hold.*

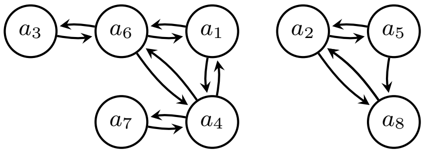
- *For every $\langle L, c \rangle \in \mathtt{Arg}(\mathcal{Y})$, the set L is consistent,*
- $L \in \bigcup\limits_{I \in \mathcal{Y}} \mathtt{g}_p(I)$ *iff* $\langle L, c \rangle \in \mathtt{Arg}(\mathcal{Y})$*,*
- *For every $I \in \mathcal{Y}$, $\exists \langle L, \mathtt{R}(I) \rangle \in \mathtt{Arg}(\mathcal{Y})$ such that $L \subseteq I$,*
- $\mathtt{cov_i}(\mathtt{Arg}(\mathcal{Y})) = \mathcal{Y}$*,*
- $\mathtt{cov_c}(\mathtt{Arg}(\mathcal{Y})) = \mathtt{C}$ *iff* $\{\mathtt{R}(I) \mid I \in \mathcal{Y}\} = \mathtt{C}$*,*
- *The set* $\mathtt{Arg}(\mathcal{Y})$ *is finite.*

Arguments may be conflicting, particularly when they violate the coherence property, i.e., their supports are consistent but their conclusions are different.

Definition 10 (Attack)**.** *Let $a = \langle L, c \rangle$, $a' = \langle L', c' \rangle \in \mathtt{Arg}(\mathcal{Y})$. We say that a* attacks *a' iff $L \cup L'$ is consistent and $c \neq c'$. We denote by $\mathtt{Att}(a)$ the set of all attackers of a.*

Property 1. *The attack relation is symmetric and irreflexive.*

Example 2 (Cont.) The attacks between the eight arguments are depicted in the figure below.



Arguments and their attack relations form an argumentation system as follows.

Definition 11 (Argumentation system)**.** *An* argumentation system *built from $\mathcal{Y}$ is a pair $AS = \langle \mathtt{Arg}(\mathcal{Y}), \mathcal{R} \rangle$ where $\mathcal{R} \subseteq \mathtt{Arg}(\mathcal{Y}) \times \mathtt{Arg}(\mathcal{Y})$ such that for $a, b \in \mathtt{Arg}(\mathcal{Y})$, $(a, b) \in \mathcal{R}$ iff a attacks b (in the sense of Def. 10).*

Since arguments are conflicting, they should be evaluated using a semantics. In this paper, we consider an extension-based semantics introduced in [Dung, 1995], namely *stable* semantics. It computes sets of arguments that can be jointly accepted. Each set is called a *stable extension* and represents a set of compatible plausible explanations.

Definition 12 (Stable Semantics)**.** *Let $AS = \langle \mathtt{Arg}(\mathcal{Y}), \mathcal{R} \rangle$ and $\mathcal{E} \subseteq \mathtt{Arg}(\mathcal{Y})$.*

- *$\mathcal{E}$ is* conflict-free *iff $\nexists a, b \in \mathcal{E}$ such that $(a, b) \in \mathcal{R}$.*
- *$\mathcal{E}$ is a* stable extension *iff it is conflict-free and $\forall a \in \mathtt{Arg}(\mathcal{Y}) \setminus \mathcal{E}$, $\exists b \in \mathcal{E}$ such that $(b, a) \in \mathcal{R}$.*

Let $\sigma(AS)$ denote the set of all stable extensions of AS.

Example 2 (Cont.) The AS depicted in the above figure has nine stable extensions.

- $\mathcal{E}_1 = \{a_1, a_2, a_3, a_7\} \quad \mathcal{E}_2 = \{a_1, a_3, a_5, a_7\}$
- $\mathcal{E}_3 = \{a_1, a_3, a_7, a_8\} \quad \mathcal{E}_4 = \{a_2, a_3, a_4\}$
- $\mathcal{E}_5 = \{a_3, a_4, a_5\} \quad \mathcal{E}_6 = \{a_3, a_4, a_8\}$
- $\mathcal{E}_7 = \{a_2, a_6, a_7\} \ \mathcal{E}_8 = \{a_5, a_6, a_7\} \ \mathcal{E}_9 = \{a_6, a_7, a_8\}$

An argumentation system has one stable extension if the attack relation is empty and multiple extensions otherwise.

Proposition 4. *Let $\mathcal{Y} \subseteq \mathbb{I}_{\mathtt{T}}$ and $AS = \langle \mathtt{Arg}(\mathcal{Y}), \mathcal{R} \rangle$.*

- $\sigma(AS) \neq \emptyset$*,*
- $\sigma(AS) = \{\mathtt{Arg}(\mathcal{Y})\}$ *iff* $\mathcal{R} = \emptyset$*.*

Let us now turn to the evaluation of individual arguments. Accepted arguments are defined in our approach using two parameters: *selection function* and *inference rule*. The former selects a subset of stable extensions and the latter selects arguments from the chosen extensions. We define various instantiations of the two parameters, capturing different *criteria* for solving conflicts between arguments.

Definition 13 (Selection Functions)**.** *Let* $\Sigma = \{\mathcal{E}_1, \ldots, \mathcal{E}_k\}$ *such that for any* $i \in \{1, \ldots, k\}$*,* $\mathcal{E}_i \subseteq \texttt{Arg}(\mathcal{Y})$*. We define below* selection functions*:*

- $\texttt{Max}(\Sigma) = \Sigma$
- $\texttt{Card}(\Sigma) = \{\mathcal{E} \in \Sigma \mid \forall \mathcal{E}' \in \Sigma, |\mathcal{E}| \geq |\mathcal{E}'|\}$
- $\texttt{Incl}_{\texttt{i}}(\Sigma) = \{\mathcal{E} \in \Sigma \mid \texttt{cov}_{\texttt{i}}(\mathcal{E})$ *is subset-maximal*$\}$
- $\texttt{Card}_{\texttt{i}}(\Sigma) = \{\mathcal{E} \in \Sigma \mid \forall \mathcal{E}' \in \Sigma,\ |\texttt{cov}_{\texttt{i}}(\mathcal{E})| \geq |\texttt{cov}_{\texttt{i}}(\mathcal{E}')|\}$
- $\texttt{Incl}_{\texttt{c}}(\Sigma) = \{\mathcal{E} \in \Sigma \mid \texttt{cov}_{\texttt{c}}(\mathcal{E})$ *is subset-maximal*$\}$
- $\texttt{Card}_{\texttt{c}}(\Sigma) = \{\mathcal{E} \in \Sigma \mid \forall \mathcal{E}' \in \Sigma, |\texttt{cov}_{\texttt{c}}(\mathcal{E})| \geq |\texttt{cov}_{\texttt{c}}(\mathcal{E}')|\}$
- $\texttt{Mix}(\Sigma) = \texttt{Card}_{\texttt{c}}(\texttt{Card}_{\texttt{i}}(\Sigma))$

Applied to the set of stable extensions of an argumentation system, the function Max returns all the extensions, Card selects the extensions that contain more arguments, the two functions $\texttt{Incl}_{\texttt{i}}, \texttt{Card}_{\texttt{i}}$ focus on the instances covered by the extensions and choose extensions with more instances. These functions promote the Success principle, which requires an explainer to have at least one explanation for each instance. The functions $\texttt{Incl}_{\texttt{c}}, \texttt{Card}_{\texttt{c}}$ focus on classes being justified by arguments. As we will se later, these principles promote explaining a large number of classes. This is useful when explanations are provided for classifier designers as they describe classifier's behaviour. Finally, the function Mix combines $\texttt{Card}_{\texttt{i}}$ and $\texttt{Card}_{\texttt{c}}$, indeed, it starts by selecting the extensions that cover more instances, then it refines the result by selecting extensions that explain more classes.

Example 2 (Cont.) Recall that $\sigma(AS) = \{\mathcal{E}_1, \ldots, \mathcal{E}_9\}$.

- $\texttt{cov}_{\texttt{i}}(\mathcal{E}_1) = \{I_2, I_3, I_4, I_5, I_6\}$
- $\texttt{cov}_{\texttt{i}}(\mathcal{E}_2) = \{I_1, I_2, I_3, I_4, I_5, I_6\}$
- $\texttt{cov}_{\texttt{i}}(\mathcal{E}_3) = \{I_2, I_3, I_4, I_5, I_6, I_7\}$ with $\texttt{cov}_{\texttt{c}}(\mathcal{E}_1) = \texttt{cov}_{\texttt{c}}(\mathcal{E}_2) = \texttt{cov}_{\texttt{c}}(\mathcal{E}_3) = \{c_0, c_1, c_2\}$
- $\texttt{cov}_{\texttt{i}}(\mathcal{E}_4) = \{I_1, I_2, I_3, I_5\}$ $\texttt{cov}_{\texttt{c}}(\mathcal{E}_4) = \{c_0, c_1\}$
- $\texttt{cov}_{\texttt{i}}(\mathcal{E}_5) = \{I_1, I_2, I_5\}$ $\texttt{cov}_{\texttt{c}}(\mathcal{E}_5) = \{c_1\}$
- $\texttt{cov}_{\texttt{i}}(\mathcal{E}_6) = \{I_1, I_2, I_5, I_7\}$ $\texttt{cov}_{\texttt{c}}(\mathcal{E}_6) = \{c_1, c_2\}$
- $\texttt{cov}_{\texttt{i}}(\mathcal{E}_7) = \{I_3, I_4, I_6, I_7\}$ $\texttt{cov}_{\texttt{c}}(\mathcal{E}_7) = \{c_0, c_2\}$
- $\texttt{cov}_{\texttt{i}}(\mathcal{E}_8) = \{I_1, I_2, I_4, I_5, I_6, I_7\}$ $\texttt{cov}_{\texttt{c}}(\mathcal{E}_8) = \{c_1, c_2\}$
- $\texttt{cov}_{\texttt{i}}(\mathcal{E}_9) = \{I_4, I_6, I_7\}$ $\texttt{cov}_{\texttt{c}}(\mathcal{E}_9) = \{c_2\}$

The selection functions return the following extensions:

- $\texttt{Card}(\sigma(AS)) = \texttt{Card}_{\texttt{c}}(\sigma(AS)) = \texttt{Incl}_{\texttt{c}}(\sigma(AS)) = \{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3\}$
- $\texttt{Incl}_{\texttt{i}}(\sigma(AS)) = \texttt{Card}_{\texttt{i}}(\sigma(AS)) = \{\mathcal{E}_2, \mathcal{E}_3, \mathcal{E}_8\}$
- $\texttt{Mix}(\sigma(AS)) = \{\mathcal{E}_2, \mathcal{E}_3\}$

We show next the links between the selection functions.

Proposition 5. *Let* Σ *be a non-empty set of subsets of arguments of* $\texttt{Arg}(\mathcal{Y})$*. The following inclusions hold:*

- $\texttt{Card}(\Sigma) \subseteq \texttt{Max}(\Sigma)$
- $\texttt{Mix}(\Sigma) \subseteq \texttt{Card}_{\texttt{i}}(\Sigma) \subseteq \texttt{Incl}_{\texttt{i}}(\Sigma) \subseteq \texttt{Max}(\Sigma)$
- $\texttt{Card}_{\texttt{c}}(\Sigma) \subseteq \texttt{Incl}_{\texttt{c}}(\Sigma) \subseteq \texttt{Max}(\Sigma)$

The selection functions may still return several extensions, hence we need to identify the strongest arguments which will yield explanations. For that purpose, we introduce two inference rules that provide strong arguments from extensions.

Definition 14 (Inference Rules)**.** *Let* Σ *be a non-empty set of subsets of arguments of* $\texttt{Arg}(\mathcal{Y})$ *and* $a \in \texttt{Arg}(\mathcal{Y})$*. We define the following* inference rules*:*

- Universal inference*:* $\Sigma \mathrel{|\!\sim}^{\forall} a$ *iff* $a \in \bigcap_{\mathcal{E} \in \Sigma} \mathcal{E}$.
- Existential inference*:* $\Sigma \mathrel{|\!\sim}^{\exists} a$ *iff* $\exists \mathcal{E} \in \Sigma$ *s.t.* $a \in \mathcal{E}$.

The next result shows the links between the two rules.

Proposition 6. *Let* Σ *be a non-empty set of subsets of arguments of* $\texttt{Arg}(\mathcal{Y})$ *and* $a \in \texttt{Arg}(\mathcal{Y})$*. The following hold:*

- $\Sigma \mathrel{|\!\sim}^{\forall} a \Rightarrow \Sigma \mathrel{|\!\sim}^{\exists} a$
- *If* $|\Sigma| = 1$*, then* $\Sigma \mathrel{|\!\sim}^{\forall} a \iff \Sigma \mathrel{|\!\sim}^{\exists} a$*.*

Selection functions and inference rules are combined for defining *accepted* arguments. Each pair gives birth to a criterion for declaring an argument as accepted.

Definition 15 (Accepted Arguments)**.** *Let* $AS = \langle \texttt{Arg}(\mathcal{Y}), \mathcal{R} \rangle$*,* α *be a selection function and* β *an inference rule. An argument* $a \in \texttt{Arg}(\mathcal{Y})$ *is* accepted*, denoted by* $AS \mathrel{|\!\sim}^{\alpha,\beta} a$*, iff* $\alpha(\sigma(AS)) \mathrel{|\!\sim}^{\beta} a$*.*

We show that accepted arguments under the function Max (which retains all extensions) are non-attacked ones if Max is combined with the universal rule and they are all arguments of the system when Max is combined with the existential rule.

Proposition 7. *Let* $AS = \langle \texttt{Arg}(\mathcal{Y}), \mathcal{R} \rangle$ *and* $a \in \texttt{Arg}(\mathcal{Y})$*.*

- $AS \mathrel{|\!\sim}^{\texttt{Max},\forall} a \iff \texttt{Att}(a) = \emptyset$
- $\{a \in \texttt{Arg}(\mathcal{Y}) \mid AS \mathrel{|\!\sim}^{\texttt{Max},\exists} a\} = \texttt{Arg}(\mathcal{Y})$

Below are links between accepted arguments returned using the same inference rule but distinct selection functions.

Proposition 8. *Let* $AS = \langle \texttt{Arg}(\mathcal{Y}), \mathcal{R} \rangle$*,* $a \in \texttt{Arg}(\mathcal{Y})$*, and* α, α' *be two selection functions. If* $\alpha(\Sigma) \subseteq \alpha'(\Sigma)$*, then:*

- $AS \mathrel{|\!\sim}^{\alpha',\forall} a \ \Rightarrow\ AS \mathrel{|\!\sim}^{\alpha,\forall} a$
- $AS \mathrel{|\!\sim}^{\alpha,\exists} a \ \Rightarrow\ AS \mathrel{|\!\sim}^{\alpha',\exists} a$

Below is a complete list of links between sets of accepted arguments returned by pairs of selection principles and inference rules.

Proposition 9. *The following implications hold.*

- $AS \mathrel{|\!\sim}^{\alpha,\forall} a \Rightarrow AS \mathrel{|\!\sim}^{\alpha,\exists} a, \forall \alpha$
- $AS \mathrel{|\!\sim}^{\texttt{Max},\forall} a \Rightarrow AS \mathrel{|\!\sim}^{\texttt{Card},\forall} a$
- $AS \mathrel{|\!\sim}^{\texttt{Max},\forall} a \Rightarrow AS \mathrel{|\!\sim}^{\texttt{Incl}_{\texttt{i}},\forall} a \Rightarrow AS \mathrel{|\!\sim}^{\texttt{Card}_{\texttt{i}},\forall} a \Rightarrow AS \mathrel{|\!\sim}^{\texttt{Mix},\forall} a$
- $AS \mathrel{|\!\sim}^{\texttt{Max},\forall} a \Rightarrow AS \mathrel{|\!\sim}^{\texttt{Incl}_{\texttt{c}},\forall} a \Rightarrow AS \mathrel{|\!\sim}^{\texttt{Card}_{\texttt{c}},\forall} a$
- $AS \mathrel{|\!\sim}^{\texttt{Card},\exists} a \Rightarrow AS \mathrel{|\!\sim}^{\texttt{Max},\exists} a$
- $AS \mathrel{|\!\sim}^{\texttt{Mix},\exists} a \Rightarrow AS \mathrel{|\!\sim}^{\texttt{Card}_{\texttt{i}},\exists} a \Rightarrow AS \mathrel{|\!\sim}^{\texttt{Incl}_{\texttt{i}},\exists} a \Rightarrow AS \mathrel{|\!\sim}^{\texttt{Max},\exists} a$

$\mathcal{Y}$	I_1	I_2	I_3	I_4	I_5	I_6	I_7
$\mathrm{g}^{\mathtt{Max},\forall}$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
$\mathrm{g}^{\mathtt{Card},\forall}$	$\emptyset$	$\{L_2\}$	$\{L_1\}$	$\{L_4\}$	$\{L_2\}$	$\{L_4\}$	$\emptyset$
$\mathrm{g}^{\mathtt{Incl_i},\forall}$	$\emptyset$	$\emptyset$	$\emptyset$	$\{L_4\}$	$\emptyset$	$\{L_4\}$	$\emptyset$
$\mathrm{g}^{\mathtt{Card_i},\forall}$	$\emptyset$	$\emptyset$	$\emptyset$	$\{L_4\}$	$\emptyset$	$\{L_4\}$	$\emptyset$
$\mathrm{g}^{\mathtt{Incl_c},\forall}$	$\emptyset$	$\{L_2\}$	$\{L_1\}$	$\{L_4\}$	$\{L_2\}$	$\{L_4\}$	$\emptyset$
$\mathrm{g}^{\mathtt{Card_c},\forall}$	$\emptyset$	$\{L_2\}$	$\{L_1\}$	$\{L_4\}$	$\{L_2\}$	$\{L_4\}$	$\emptyset$
$\mathrm{g}^{\mathtt{Mix},\forall}$	$\emptyset$	$\{L_2\}$	$\{L_1\}$	$\{L_4\}$	$\{L_2\}$	$\{L_4\}$	$\emptyset$

Table 1: The outcomes of all functions $\mathrm{g}^{\alpha,\forall}$ in Example 2.

- $AS \mathrel{|\!\sim}^{\mathtt{Card_c},\exists} a \Rightarrow AS \mathrel{|\!\sim}^{\mathtt{Incl_c},\exists} a \Rightarrow AS \mathrel{|\!\sim}^{\mathtt{Max},\exists} a$

We are now ready to define our new parameterized family of plausible explanation functions. For a given instance I, they return the support of any argument in favour of $\mathtt{R}(I)$ inferred by following one of the principles defined above. The support of the argument should be part of the instance I.

Definition 16 (Explanation Functions)**.** *Let* $\mathtt{T}$ *be a theory,* $\mathcal{Y} \subseteq \mathbb{I}_{\mathtt{T}}$, $\mathtt{R}$ *a classifier,* α *a selection function and* β *an inference rule. An explainer is a function* $\mathrm{g}^{\alpha,\beta}$ *mapping every instance* $I \in \mathcal{Y}$ *into a set of subsets of literals such that every* $L \in \mathrm{g}^{\alpha,\beta}(I)$ *satisfies the following:*

- $AS \mathrel{|\!\sim}^{\alpha,\beta} \langle L, \mathtt{R}(I)\rangle$ *where* $AS = \langle \mathtt{Arg}(\mathcal{Y}), \mathcal{R}\rangle$*,*
- $L \subseteq I$*.*

Example 2 (Cont.) Table 1 summarizes the explanations of the seven instances provided by every function which uses the universal inference rule. Note that the new functions explain more instances than the argument-based function (which is equivalent to $\mathrm{g}^{\mathtt{Max},\forall}$) from [Amgoud, 2021b].

We show that all the above defined explanation functions are refined plausible explainers, i.e., they return subsets of explanations computed by the function g_p (see Definition 8).

Proposition 10. *Let* $\mathtt{T}$ *be a theory,* $\mathcal{Y} \subseteq \mathbb{I}_{\mathtt{T}}$ *and* $\mathtt{R}$ *a classifier. For every selection function* α*, every inference rule* β*, every* $I \in \mathcal{Y}$*, it holds that* $\mathrm{g}^{\alpha,\beta}(I) \subseteq \mathrm{g}_p(I)$*.*

The following results show the links between the various explanation functions.

Proposition 11. *Let* $I \in \mathbb{I}_{\mathtt{T}}$*.*

- $\mathrm{g}^{\alpha,\forall}(I) \subseteq \mathrm{g}^{\alpha,\exists}(I)$ *for any selection function* α
- $g^{\mathtt{Max},\forall}(I) \subseteq g^{\mathtt{Card},\forall}(I)$
- $g^{\mathtt{Max},\forall}(I) \subseteq g^{\mathtt{Incl_i},\forall}(I) \subseteq g^{\mathtt{Card_i},\forall}(I) \subseteq g^{\mathtt{Mix},\forall}(I)$
- $g^{\mathtt{Max},\forall}(I) \subseteq g^{\mathtt{Incl_c},\forall}(I) \subseteq g^{\mathtt{Card_c},\forall}(I)$
- $g^{\mathtt{Card},\exists}(I) \subseteq g^{\mathtt{Max},\exists}(I)$
- $g^{\mathtt{Mix},\exists}(I) \subseteq g^{\mathtt{Card_i},\exists}(I) \subseteq g^{\mathtt{Incl_i},\exists}(I) \subseteq g^{\mathtt{Max},\exists}(I)$
- $g^{\mathtt{Card_c},\exists}(I) \subseteq g^{\mathtt{Incl_c},\exists}(I) \subseteq g^{\mathtt{Max},\exists}(I)$

It is worth mentioning that the explanation function $\mathrm{g}^{\mathtt{Max},\exists}$ corresponds exactly to the plausible explanation function g_p.

Property 2. *It holds that* $\mathrm{g}^{\mathtt{Max},\exists} = \mathrm{g}_p$*.*

The function $\mathrm{g}^{\mathtt{Max},\forall}$ coincides with the function g^* introduced in [Amgoud, 2021b]. We show that this function is very cautious as it discards any explanation which is incoherent with at least one other explanation.

Proposition 12. *Let* $AS = \langle \mathtt{Arg}(\mathcal{Y}), \mathcal{R}\rangle$ *and* $I \in \mathcal{Y}$*.* $\mathrm{g}^{\mathtt{Max},\forall}(I) = \{L \in \mathrm{g}_p(I) \mid \forall L' \in \mathrm{g}_p(I')$*, if* $\mathtt{R}(I) \neq \mathtt{R}(I')$ *then* $L \cup L'$ *is inconsistent*$\}$*.*

We show next how the various explainers behave wrt the two principles of Coherence and Success.

Theorem 2. *Let* $AS = \langle \mathtt{Arg}(\mathcal{Y}), \mathcal{R}\rangle$*,* α *be a selection function and* β *an inference rule. If* $|\sigma(AS)| > 1$*, then:*

- $\mathrm{g}^{\alpha,\beta}$ *satisfies Success iff* $\alpha = \mathtt{Max}$ *and* $\beta = \exists$ *.*
- $\mathrm{g}^{\alpha,\beta}$ *satisfies Coherence iff* $\beta = \forall$*.*

The above result shows that $\mathrm{g}^{\mathtt{Max},\exists}$ (or g_p) is the **only** function which satisfies success and all the other functions that are based on the existential inference rule violate both success and coherence. Consequently, those functions are not reasonable. Note that in this paper, we investigated the different possibilities for the purpose of *completeness* and proving formally which function is not suitable and why it is not.

Coherence is guaranteed by **all the functions that are based on the universal inference rule**. Furthermore, $\mathrm{g}^{\mathtt{Card},\forall}$, $\mathrm{g}^{\mathtt{Mix},\forall}$ and $\mathrm{g}^{\mathtt{Card_c},\forall}$ are more informative than the other functions that use the same inference rule. Indeed, they provide more explanations for instances, and can thus **explain more instances**. However, the three functions may return different outcomes as they follow **different strategies**. $\mathrm{g}^{\mathtt{Card},\forall}$ is less interesting than the two others. It selects the extensions that contain more arguments, but the latter may support the same class as in our running example (the arguments in the extension $\mathcal{E}_1$ are all in favour of the class c_0). Hence, any instance whose prediction is c_1 gets an empty set of explanations.

The function $\mathrm{g}^{\mathtt{Mix},\forall}$ **maximizes the number of instances** for which explanations are provided, this is important in domains like healthcare or banking where explanations are generally requested by end-users.

$\mathrm{g}^{\mathtt{Card_c},\forall}$ **maximizes the number of explained classes**. It is suitable for understanding the global behaviour of a classifier, especially for a problem with a lot of classes.

6 Experimental Analysis

Recall that the function g_p generates from a sample all possible abductive explanations, which may be **incorrect**. Our novel functions that use the universal inference rule guarantee correctness of explanations at the **cost of success**. The aim of this section is to confirm experimentally these findings. We provide two experiments, the first one measures the proportion of correct explanations provided by g_p while the second measures the proportion of explained instances by novel functions that satisfy coherence. We implemented five functions: g_p, the argument-based function $g^{\mathtt{Max},\forall}$ from [Amgoud, 2021b] and the three new functions that follow different strategies: $g^{\mathtt{Card},\forall}$, $g^{\mathtt{Card_i},\forall}$ and $g^{\mathtt{Card_c},\forall}$. Details on the implementation are given in the supplementary material.

We tested the four functions on various datasets, namely *diabetes*, *titanic* available on the *Kaggle* website, and *lending adult* (shortened) and *recidivism* (shortened) that are available on Anchors' experiments [Ribeiro *et al.*, 2018]. In each experiment, we built the whole feature space, shuffled the instances, and kept different percentages of the whole space from which we generate arguments.

$\%space$	adult	diabetes	lending	titanic	rcdv
12.50	0.00	0.00	0.00	0.00	0.00
25.00	0.00	0.01	0.02	0.00	0.00
37.50	0.05	0.19	0.14	1.77	0.00
50.00	0.52	1.01	0.85	3.48	0.04
62.50	2.21	4.12	3.98	8.14	0.39
75.00	7.78	13.18	13.76	16.39	2.84
87.50	27.36	37.60	41.51	32.63	16.03
93.75	50.48	60.36	66.66	50.74	38.96
100.00	100.00	100.00	100.00	100.00	100.00

Table 2: Percentage of unattacked arguments computed from the portion *% space* of the whole feature space.

For checking the proportion of correct explanations provided by g_p in a dataset, we measure the percentage of non-attacked arguments generated from the dataset. Table 2 shows that it is very unlikely to find correct explanations by probing a model with a small part of the feature space. This might not entirely reflect the nature of a problem if natural restrictions of possible values confine real instances to a small subspace (something usually hard to know in practice).

Function	adult	diabetes	titanic	rcdv	lending
$g^{\mathtt{Max},\forall}$	0	0	0	0	0
$g^{\mathtt{Card_c},\forall}$	0	0	0	0	0
$g^{\mathtt{Card},\forall}$	40	60	43	48	48
$g^{\mathtt{Card_i},\forall}$	14	36	81	96	42

Table 3: Percentage of explained instances for each dataset.

Table 3 gives the proportion of instances of a dataset that are explained by the argument-based function $g^{\mathtt{Max},\forall}$ and the three new functions, given the guarantee of coherence according to Theorem 2. While $g^{\mathtt{Max},\forall}$ fails to explain any instance in the chosen datasets, $g^{\mathtt{Card},\forall}$ and $g^{\mathtt{Card_i},\forall}$ show better performances. $g^{\mathtt{Card_c},\forall}$, which maximises the number of explained **classes**, fails as well but for the simple reason that the classification theories of the datasets contain only two classes.

7 Related Work

Most work on finding explanations in the ML literature is experimental, focusing on specific models, exposing their internal representations to find correlations *post hoc* between these representations and the predictions. There haven't been a lot of formal characterizations of explanations in AI, with the exception of [Ignatiev *et al.*, 2019b], which defines abductive explanations and adversarial examples in a fragment of first order logic, [Darwiche and Hirth, 2020], who focused on semi-factuals, that they consider a specific form of counterfactuals, [Amgoud, 2021a] who defined in a unified setting abductive explanations, counterfactuals and contrastive explanations. These works generate explanations from the whole feature space, which in practice is not reasonable. In our work, we generate abductive explanations for black-box classifiers from subsets of instances. Unlike existing approaches which are based on samples, our novel functions that use the universal inference rule offer theoretical guarantees as they ensure correctness of explanations. They have also a greater explanatory power than the unique function in the literature that satisfies correctness, namely the argument-based function from [Amgoud, 2021b].

Unlike our work which explains existing black-box models, [Cocarascu *et al.*, 2020] proposed classification models that are based on arguments. Their explanations are defined in dialectical way as fictitious dialogues between a proponent (supporting an output) and an opponent (attacking the output) following [Dung, 1995]. The authors in [Zhong *et al.*, 2019; Rago *et al.*, 2018] followed the same approach for defining explainable multiple decision systems or scheduling systems.

In [Borg and Bex, 2021; Liao and van der Torre, 2020], the authors investigated explainability of argument status in Dung style argumentation setting. This is quite far from our work which uses argumentation for explaining ML models.

8 Conclusion

The paper proposed an argumentation-based approach for defining explanation functions for black-box classifiers. The novel functions provide abductive explanations from samples of instances, some of them guarantee correctness of their outcomes while one function ensures existence of explanations.

This work lends itself to a number of developments in order to improve its generality and the compromise coherence/success of the explanations. We can consider other attack relation for minimizing the distance between plausible and absolute explanations. The idea is to take into account weights of features and thus assign a basic weight to every argument reflecting the importance of features it is based on.

Acknowledgements

This work was supported by the ANR (ANR-19-PI3A-0004) through the AI Interdisciplinary Institute ANITI and by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

References

[Amgoud and Ben-Naim, 2022] Leila Amgoud and Jonathan Ben-Naim. Axiomatic foundations of explainability. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 636–642, 2022.

[Amgoud, 2021a] Leila Amgoud. Explaining black-box classification models with arguments. In *33rd IEEE International Conference on Tools with Artificial Intelligence, ICTAI*, pages 791–795, 2021.

[Amgoud, 2021b] Leila Amgoud. Non-monotonic explanation functions. In *Proceedings of the 16th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, ECSQARU*, pages 19–31, 2021.

[Audemard *et al.*, 2022] Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. On preferred abductive explanations for decision trees and random forests. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 643–650, 2022.

[Borg and Bex, 2021] AnneMarie Borg and Floris Bex. Necessary and sufficient explanations for argumentation-based conclusions. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 16th European Conference*, pages 45–58, 2021.

[Burkart and Huber, 2021] Nadia Burkart and Marco F. Huber. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317, 2021.

[Cocarascu *et al.*, 2020] Oana Cocarascu, Andria Stylianou, Kristijonas Cyras, and Francesca Toni. Data-empowered argumentation for dialectically explainable predictions. In *24th European Conference on Artificial Intelligence ECAI*, volume 325, pages 2449–2456. IOS Press, 2020.

[Cooper and Marques-Silva, 2021] Martin C. Cooper and João Marques-Silva. On the tractability of explaining decisions of classifiers. In *CP 2021*, pages 21:1–21:18, 2021.

[Cyras *et al.*, 2019a] Kristijonas Cyras, David Birch, Yike Guo, Francesca Toni, Rajvinder Dulay, Sally Turvey, Daniel Greenberg, and Tharindi Hapuarachchi. Explanations by arbitrated argumentative dispute. *Expert Systems with Applications*, 127:141–156, 2019.

[Cyras *et al.*, 2019b] Kristijonas Cyras, Dimitrios Letsios, Ruth Misener, and Francesca Toni. Argumentation for explainable scheduling. In *The Thirty-Third Conference on Artificial Intelligence, AAAI*, pages 2752–2759, 2019.

[Darwiche and Hirth, 2020] Adnan Darwiche and Auguste Hirth. On the reasons behind decisions. In *24th European Conference on Artificial Intelligence ECAI*, volume 325, pages 712–720. IOS Press, 2020.

[Dimopoulos *et al.*, 1997] Yannis Dimopoulos, Saso Dzeroski, and Antonis Kakas. Integrating explanatory and descriptive learning in ILP. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, IJCAI*, pages 900–907, 1997.

[Dung, 1995] Phan Minh Dung. On the Acceptability of Arguments and its Fundamental Role in Non-Monotonic Reasoning, Logic Programming and n-Person Games. *Artificial Intelligence*, 77:321–357, 1995.

[Ignatiev *et al.*, 2019a] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. Abduction-based explanations for machine learning models. In *National Conference on Artificial Intelligence*, pages 1511–1519, 2019.

[Ignatiev *et al.*, 2019b] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. On relating explanations and adversarial examples. In *Conference on Neural Information Processing Systems*, pages 15857–15867, 2019.

[Kakas and Riguzzi, 2000] Antonis Kakas and Fabrizio Riguzzi. Abductive concept learning. *New Generation Computing*, 18(3):243–294, 2000.

[Liao and van der Torre, 2020] Beishui Liao and Leendert van der Torre. Explanation semantics for abstract argumentation. In *Computational Models of Argument*, volume 326 of *Frontiers in Artificial Intelligence and Applications*, pages 271–282. IOS Press, 2020.

[Narodytska *et al.*, 2019] Nina Narodytska, Aditya A. Shrotri, Kuldeep S. Meel, Alexey Ignatiev, and João Marques-Silva. Assessing heuristic machine learning explanations with model counting. In *22nd International Conference on Theory and Applications of Satisfiability Testing*, pages 267–278, 2019.

[Rago *et al.*, 2018] Antonio Rago, Oana Cocarascu, and Francesca Toni. Argumentation-based recommendations: Fantastic explanations and how to find them. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI*, pages 1949–1955, 2018.

[Rahwan and Simari, 2009] Iyad Rahwan and Guillermo Simari, editors. *Argumentation in Artificial Intelligence*. Springer, 2009.

[Ribeiro *et al.*, 2016] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should itrust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1135–1144, 2016.

[Ribeiro *et al.*, 2018] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the Thirty-Second Conference on Artificial Intelligence*, pages 1527–1535, 2018.

[Shih *et al.*, 2018] Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining Bayesian network classifiers. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI*, pages 5103–5111, 2018.

[Zhong *et al.*, 2019] Qiaoting Zhong, Xiuyi Fan, Xudong Luo, and Francesca Toni. An explainable multi-attribute decision model based on argumentation. *Expert Systems with Applications*, 117:42–61, 2019.