Argument-based Explanation Functions

Leila Amgoud, Philippe Muller, Henri Trenquier

Argument-based Explanation Functions

Extended Abstract

Leila Amgoud
CNRS – IRIT
Toulouse, France
leila.amgoud@irit.fr

Philippe Muller
University of Toulouse – IRIT
Toulouse, France
philippe.muller@irit.fr

Henri Trenquier
University of Toulouse – ANITI
Toulouse, France
henri.trenquier@univ-tlse3.fr

ABSTRACT

Explaining predictions made by inductive classifiers whose internal reasoning is left unspecified (black-boxes) is becoming a hot topic. *Abductive explanations* are one of the most popular types of explanations that are provided for the purpose. They are sufficient reasons for making predictions. They are generated from the whole feature space, which is not reasonable in practice. This paper investigates functions that generate abductive explanations from a set of instances. It shows that such explainers 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 general argumentation-based setting in which various functions satisfying one of the two properties are defined.

KEYWORDS

Classification, Explainability, Argumentation.

ACM Reference Format:
Leila Amgoud, Philippe Muller, and Henri Trenquier. 2023. Argument-based Explanation Functions: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Explaining predictions of black-box classification models has become a vital need, and has generated a lot of research (see [6, 7, 9, 12, 14, 15, 21] for more on explainability). 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. Such explanations are generally generated from the whole feature space (eg., [1, 4, 10, 13]), which is reasonable when models are interpretable, like Decision Trees, but not tractable in case of black-boxes [8]. As a solution, the two prominent explanation functions Anchors [19] and LIME [18] and the argument-based function [2] generate abductive explanations from a sample (i.e., subset) of instances, avoiding thus exploring the whole feature space. However, it has been shown in [2, 16] that 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. It is also very cautious as it discards all conflicting explanations generated from the sample.

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

This paper investigates explanation functions that generate abductive explanations from a sample while satisfying desirable properties. It starts by proving an impossibility result, which states that a function which generates abductive explanations from a sample cannot guarantee both existence of explanations (success) and their correctness (coherence). Then, it proposes a parametrized argumentation-based approach for defining in a systematic way various functions satisfying one of the two incompatible properties.

2 PRELIMINARIES

Throughout the paper, we consider a *classification theory* as a tuple $\mathtt{T} = \langle \mathtt{F}, \mathtt{dom}, \mathtt{C} \rangle$ made of a finite set $\mathtt{F}$ of *features*, a function $\mathtt{dom}$ which returns the *domain* of every feature and a finite set $\mathtt{C}$ of *classes*. We call *literal* any pair (f, v) where $f \in \mathtt{F}$ and $v \in \mathtt{dom}(f)$ and *instance* any subset of literals in which every attribute $f \in \mathtt{F}$ appears exactly once. We denote by $\mathtt{Lit}(\mathtt{T})$ the set of all possible literals of a theory $\mathtt{T}$ and by $\mathtt{Inst}(\mathtt{T})$ the set of all its instances, called also *feature space*. We say that a set of literals is *consistent* if it does not contain two literals having the same feature but distinct values. A *classification model*, or *classifier*, is a surjective function $\mathtt{R}$ that assigns a single class from $\mathtt{C}$ to every instance $I \in \mathtt{Inst}(\mathtt{T})$. An *explainer* is a function which returns the reasons (explanations) behind predicting the class of a given instance by a classifier. Throughout the paper, we assume an arbitrary theory $\mathtt{T}$ and a classifier $\mathtt{R}$.

3 ABDUCTIVE EXPLANATIONS

Abductive explanations are sufficient reasons for making a prediction. They are generated by exploring the whole feature space; we call such explanations *absolute abductive explanations*.

Definition 1. *Let $I \in \mathtt{Inst}(\mathtt{T})$. An* absolute abductive explanation *of $\mathtt{R}(I)$ is a set $L \subseteq \mathtt{Lit}(\mathtt{T})$ such that:*

- $L \subseteq I$,
- $\forall I' \in \mathtt{Inst}(\mathtt{T}) \setminus \{I\}$ *such that* $L \subseteq I'$, $\mathtt{R}(I') = \mathtt{R}(I)$,
- $\nexists L' \subset L$ *such that* L' *satisfies the above conditions.*

We denote by $\mathtt{g}_a$ the function which assigns to every instance $I \in \mathtt{Inst}(\mathtt{T})$ the set of all absolute abductive explanations of $\mathtt{R}(I)$.

Example 1. *Consider a theory made of two binary features f_1, f_2 and two classes (0,1). Assume the classifier $\mathtt{R}$ such that for $I \in \mathtt{Inst}(\mathtt{T})$, $\mathtt{R}(I) = f_1 \vee f_2$. For $I = \{(f_1, 1), (f_2, 1)\}$, $\mathtt{R}(I) = 1$ and $\mathtt{g}_a(I) = \{L_1, L_2\}$ with $L_1 = \{(f_1, 1)\}$ and $L_2 = \{(f_2, 1)\}$.*

Generating absolute explanations from the whole feature space (second condition) may not be feasible in practice especially for black-box classifiers. Hence, we introduce *plausible abductive explanations*, which are based on a subset of instances only.

Definition 2. *Let $\mathcal{Y} \subseteq \mathtt{Inst}(\mathtt{T})$ and $I \in \mathcal{Y}$. A* plausible abductive explanation *of $\mathtt{R}(I)$ is a set $L \subseteq \mathtt{Lit}(\mathtt{T})$ such that:*

- $L \subseteq I$,

- $\forall I' \in \mathcal{Y} \setminus \{I\}$ *such that* $L \subseteq I'$, $\mathsf{R}(I') = \mathsf{R}(I)$,
- $\nexists L' \subset L$ *such that* L' *satisfies the above conditions.*

We denote by g_p *the function generating them for every instance.*

Example 2. *Assume a classification problem of deciding whether to go hiking (1) or not (0). The decision is based on four binary features: Being on vacation (V), having a concert (C), having a meeting (M) and having an exhibition (E). Assume a classifier* R *that assigns classes to instances of* $\mathcal{Y} \subset \mathtt{Inst(T)}$ *as shown in the table below.*

$\mathcal{Y}$	V	C	M	E	$\mathsf{R}(I_i)$
I_1	0	0	1	0	0
I_2	1	0	0	0	1
I_3	0	0	1	1	0
I_4	1	0	0	1	1
I_5	0	1	1	0	0
I_6	0	1	1	1	0
I_7	1	1	0	1	1

$L_1 = \{(V, 0)\}$
$L_2 = \{(M, 1)\}$
$L_3 = \{(C, 1), (E, 0)\}$
$L_4 = \{(V, 1)\}$
$L_5 = \{(M, 0)\}$

$\mathsf{g}_p(I_1) = \mathsf{g}_p(I_3) = \mathsf{g}_p(I_6) = \{L_1, L_2\}$
$\mathsf{g}_p(I_2) = \mathsf{g}_p(I_4) = \mathsf{g}_p(I_7) = \{L_4, L_5\}$
$\mathsf{g}_p(I_5) = \{L_1, L_2, L_3\}$

We show that every absolute explanation of an instance is a superset of a plausible explanation of the same instance. This shows that a plausible explanation is not larger than an absolute one.

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

The following example shows that a plausible explanation may not be the subset of any absolute explanation.

Example 2 (Cont.) Assume the instance I_8 below is labelled 1.

	V	C	M	E	$\mathsf{R}(I_8)$
I_8	1	1	0	0	1

While $L_3 \in \mathsf{g}_p(I_5)$ in $\mathcal{Y}$, L_3 cannot be (a subset of) an absolute explanation of the decision $\mathsf{R}(I_5)$.

4 IMPOSSIBILITY RESULT

A property that should be satisfied by any explainer has been introduced in [2]. It states that two explanations of instances labelled with different classes should be inconsistent. This property prevents the following three undesirable situations: Assume two instances $I, I' \in \mathtt{Inst(T)}$ such that $\mathcal{R}(I) \neq \mathcal{R}(I')$. Assume also that L is an explanation for I and L' is an explanation for I'. We may have the three cases: i) $L = L'$, ii) $L \subset L'$, or 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 $\mathcal{R}(I)$ and $\mathcal{R}(I')$ 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 predicts acceptance while the second predicts rejection of the loan.

Principle 1. *(Coherence) An explainer* g *satisfies* coherence *iff the following holds: for any classifier* R, *for any theory* T, *for all* $I, I' \in \mathtt{Inst(T)}$, *if* $\mathsf{R}(I) \neq \mathsf{R}(I')$, *then for every* $L \in \mathsf{g}(I)$, *for every* $L' \in \mathsf{g}(I')$, *we have that* $L \cup L'$ *is inconsistent.*

We introduce another property stating that an explainer should always provide outcomes. A similar property has been defined in [3] for functions that explain classes instead of instances.

Principle 2. *(Success) An explainer* g *satisfies* success *iff, for any classifier* R, *for any theory* T, *for any* $I \in \mathtt{Inst(T)}$, $\mathsf{g}(I) \neq \emptyset$.

It has been shown in [2] that g_a satisfies both properties while the function g_p satisfies Success but violates Coherence.

Example 2 (Cont.) Consider the two instances I_1 and I_2. Note that $\mathsf{R}(I_1) \neq \mathsf{R}(I_2)$ while $L_1 \in \mathsf{g}_p(I_1)$, $L_5 \in \mathsf{g}_p(I_2)$ and $L_1 \cup L_5$ is consistent. Consequently, there exists $I' \in \mathtt{Inst(T)}$ such that $L_1 \cup L_5 \subseteq I'$. Since I' is assigned a single class, then at least one of the two explanations (L_1, L_5) is incorrect.

The aim is to define explanation functions that generate plausible explanations (from samples) and that satisfy the two principles. Let us first introduce the notion of *refined plausible explainer*.

Definition 3. *Let* $\mathcal{Y} \subseteq \mathtt{Inst(T)}$. *A* refined plausible explainer *is a function* g *mapping every* $I \in \mathcal{Y}$ *into* $\mathsf{g}(I) \subseteq \mathsf{g}_p(I)$.

We show that the two principles are *incompatible* as there is no refined plausible explainer that can satisfy the two principles at the same time for every classifier, every theory, and every sample.

Theorem 1. *There is no refined plausible explainer that satisfies both Coherence and Success.*

This negative result shows that generating abductive explanations from a subset of feature space is a tricky problem and one has to choose between the quality of explanations and their existence.

5 ARGUMENT-BASED EXPLAINERS

Argumentation is a powerful approach for reasoning with conflicting or incomplete information (see [5, 17, 20] for more information). We propose a parametrized 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; an argument is a pair $\langle L, c\rangle$ where L is a plausible explanation for the class c. In Example 2, the pairs $a = \langle L_1, 0\rangle$ and $b = \langle L_5, 1\rangle$ are arguments. The approach identifies conflicts among arguments, where a conflict occurs when two arguments violate Coherence. For instance, the two arguments a and b are conflicting since L_1 and L_5 violate Coherence, hence we say that a and b attack each other. For evaluating arguments, an extension semantics from [11], namely *stable*, is used. The outcome is a set of sets of arguments that can be jointly accepted. Finally, the approach identifies *accepted arguments*, and uses the latter for defining 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. For instance, one may consider all extensions, or those that contain more arguments, or that cover more instances. The latter selects (accepted) arguments from the chosen extensions. Two criteria are investigated: the universal criterion which selects arguments belonging to all extensions and the existential one which elects every argument appearing in at least one extension. The supports of selected arguments are used for explaining instances. We define various functions combining different instantiations of the two parameters.

ACKNOWLEDGMENTS

This work was supported by the ANR (ANR-19-PI3A-0004) through the AI Interdisciplinary Institute, ANITI, as a part of France's "Investing for the Future — PIA3" program.

REFERENCES

[1] Leila Amgoud. 2021. Explaining Black-box Classification Models with Arguments. In *33rd IEEE International Conference on Tools with Artificial Intelligence, ICTAI*. 791–795.

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

[3] Leila Amgoud and Jonathan Ben-Naim. 2022. Axiomatic Foundations of Explainability. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, Luc De Raedt (Ed.). 636–642.

[4] Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. 2022. On Preferred Abductive Explanations for Decision Trees and Random Forests. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*. 643–650.

[5] Pietro Baroni, Dov Gabbay, Massimiliano Giacomin, and Leon van der Torre (Eds.). 2018. *Handbook of Formal Argumentation, Volume 1*. College Publications.

[6] Or Biran and Courtenay Cotton. 2017. Explanation and Justification in Machine Learning: A Survey. In *IJCAI Workshop on Explainable Artificial Intelligence (XAI)*. 1–6.

[7] Nadia Burkart and Marco Huber. 2021. A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research* 70 (2021), 245–317.

[8] Martin C. Cooper and João Marques-Silva. 2021. On the Tractability of Explaining Decisions of Classifiers. In *CP 2021*. 21:1–21:18.

[9] Kristijonas Cyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. 2021. Argumentative XAI: A Survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI*. 4392–4399.

[10] Adnan Darwiche and Auguste Hirth. 2020. On the Reasons Behind Decisions. In *24th European Conference on Artificial Intelligence ECAI*, Vol. 325. IOS Press, 712–720.

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

[12] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A Survey of Methods for Explaining Black Box Models. *Comput. Surveys* 51, 5 (2019), 93:1–93:42.

[13] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. 2019. On Relating Explanations and Adversarial Examples. In *Thirty-third Conference on Neural Information Processing Systems, NeurIPS*. 15857–15867.

[14] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[15] C. Molnar. 2020. *Interpretable Machine Learning*. Lulu.com. https://books.google.fr/books?id=RHjTxgEACAAJ

[16] Nina Narodytska, Aditya A. Shrotri, Kuldeep S. Meel, Alexey Ignatiev, and João Marques-Silva. 2019. Assessing Heuristic Machine Learning Explanations with Model Counting. In *22nd International Conference on Theory and Applications of Satisfiability Testing - SAT*. 267–278.

[17] Iyad Rahwan and Guillermo Simari (Eds.). 2009. *Argumentation in Artificial Intelligence*. Springer.

[18] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. 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*. 1135–1144.

[19] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*. 1527–1535.

[20] Guillermo Simari, Massimiliano Giacomin, Dov Gabbay, and Matthias Thimm (Eds.). 2021. *Handbook of Formal Argumentation, Volume 2*. College Publications.

[21] Ilia Stepin, José Maria Alonso, Alejandro Catalá, and Martin Pereira-Fariña. 2021. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access* 9 (2021), 11974–12001.