Recherches en IA explicable au sein du département IA de l'IRIT

Pascale Zaraté, Nathalie Aussenac-Gilles

▶ **To cite this version:**

Pascale Zaraté, Nathalie Aussenac-Gilles. Recherches en IA explicable au sein du département IA de l'IRIT. Bulletin de l'Association Française pour l'Intelligence Artificielle, 2022, 116, pp.15-20. hal-04726302

HAL Id: hal-04726302
https://hal.science/hal-04726302v1

Submitted on 9 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Recherches en IA explicable au sein du département IA de l'IRIT

Par

Pascal ZARATE
IRIT / ADRIA
Université Toulouse 1 Capitole
pascal.zarate@irit.fr
www.irit.fr/departement/intelligence-artificielle/adria/

Nathalie AUSSENAC
IRIT / MELODI
CNRS
nathalie.aussenac@irit.fr
www.irit.fr/departement/intelligence-artificielle/melodi/

The IRIT-Artificial Intelligence Department investigates the automation of reasoning and decision-making processes, based on knowledge drawn from texts and data, but also at defining natural language analysis systems, with a view to helping humans. This research addresses the following issues :

- Automated reasoning, especially under uncertainty and probabilistic reasoning ;
- Symbolic and statistical machine learning ;
- Decision support systems for an individual or a group of decision makers and automated decision processes ;
- The formalization of interaction and communication between agents, in particular the role of beliefs and the management of arguments ;
- The security of information and communication systems ;
- Models and methods for natural language processing, natural language semantics and discourse analysis ;
- Knowledge engineering and formal ontology, from knowledge extraction, its modelling and its formal representation, its linking within the semantic web and the web of data, and the study of its evolution.

The AI department is composed by three teams : ADRIA, LILaC and MELODI. The interactions among the 3 teams are important, with several co-supervised PhD theses and joint projects.

Explainability is addressed by a lot of researchers using different approaches :

- Formal Explainability (cf Marques-Silva and co),
- Analogical explanations (cf Prade and Richard),
- Abstract argumentation (cf Duchatelle et al.),
- Formal Reasoning for Reinforcement learning (cf Saulières et al.),
- Explainable AI for Intrusion detection (cf Chevalier),
- Interacting a machine Learning system with an explicit reasoning system : Application on medical data (cf Mayouf et al.).

Formal explainability

Joao Marques-Silva, Martin Cooper, Xuanxiang Huang, Yacine Izza, Nicholas Asher

Since 2019, our team has been investigating formal approaches to explainability in machine learning (ML), which we refer to as Formal Explainable AI (FXAI). In contrast to most of the existing work on explainability in ML, we have proposed definitions of explana-

tions that are rigorous, that take into account the underlying ML model, and that are amenable to exact computation using automated reasoners. The team currently includes João MARQUES-SILVA (CNRS DR and ANITI Research Chair), Martin COOPER (UPS Professor and ANITI Co-Chair), Yacine IZZA (Postdoctoral researcher, ANITI and IRIT), Xuanxiang HUANG (PhD student, ANITI and IRIT), Thomas GERSPACHER (former PhD student, ANITI and IRIT), and Nicholas ASHER (CNRS DR, and ANITI Scientific Director). The initial ideas on formal explainability we presented in the following papers : [10], [11] and [9]. A recent overview of the progress in formal approaches to explainability is given in [16].

Furthermore, we have demonstrated a number of results, organized as follows :

1. Tractable explainability : We have shown that, for several well-known families of classifiers, the computation of one explanation is poly-time. This is the case of Naive Bayes Classifiers (see [14]), monotonic classifiers (see [15]), decision trees and other graph-based classifiers (see [6]), and several families of propositional languages (see [24]). The tractability of several other families of classifiers is investigated in [4].
2. Connections between fairness and explainability : some initial results were presented in [7] and more recently in [2].
3. Duality of explanations : two kinds of minimal-hitting set duality relationships were identified (see [11] and [9]).
4. Practical efficient explainability : We have shown that for decision lists and sets and for tree ensembles, the computation of one explanation has been shown to be computationally hard for decision lists and sets (see [12]), random forests (see [13]) and tree ensembles in general (see [8]). However, we also developed logic encodings that enable the efficient practical computation of explanations.
5. Assessment of model-agnostic explainers : our results demonstrate the inadequacy of well-known model-agnostic explainers in settings where the rigor of explanations is paramount (see [20]).
6. Improvements to model-agnostic explainers (see [1]).
7. Trade-offs between rigor of explanations and their size : ongoing work.

Analogical explanations

Henri Prade, Gilles Richard

The approach [21] relies on the use of analogical proportions (AP), which are statements relating four items, of the form "a is to b as c is to d". The items are represented by vectors of Boolean or categorical attribute values. a, b, c, d make a valid AP, if the attributes can be split into three subsets $\mathcal{A}, \mathcal{A}', \mathcal{A}''$ (some may be empty), in such a way that a, b, c, d are identical on $\mathcal{A}$, $a = b$ and $c = d$ on $\mathcal{A}'$, while on $\mathcal{A}''$ the same change of values takes place from a to b, and from c to d. It is pictured in the table below, where s, t, u, v, w are sub-vectors of attribute values. The change of class from x to y in pair (a, b) can be explained only by the change of values of attributes in $\mathcal{A}''$. The same change for pair (c, d) has the same effect for the classes. Thus, this provides a basis for predicting or for explaining why d is in class y. Each pair may be viewed as a potential rule expressing that in a context (described by values on $\mathcal{A} \cup \mathcal{A}'$) the change from v to w induces the flip from class x to class y. The Confidence in the rule can be evaluated on the set of examples at hand. As can be seen, the approach does not require to know how the class of d has been obtained for explaining it.

	$\mathcal{A}$ full id.	$\mathcal{A}'$ pair id.	$\mathcal{A}''$ change	class
a	s	t	v	x
b	s	t	w	y
c	s	u	v	x
d	s	u	w	?

A Query-based Explanation Model for Abstract Argumentation

Théo Duchatelle, Philippe Besnard, Sylvie Doutre, Marie-Christine Lagasquie

Abstract Argumentation [5] is a rising formalism for computing explanations [22]. An approach to explain this formalism itself is pictured in Figure 1.1.

The approach includes a formal grammar for modelling the questions the user can ask, and a process for building the answers which uses graph operations and which exploits elements of the question.

FR4RL : Formal Reasoning for Reinforcement Learning

Leo Saulieres, Martin Cooper, Florence Dupin de Saint-Cyr, Joao Marques-Silva

The PhD started in October 2021. The proposed research project is positioned at the intersection of automated reasoning (AR) and Reinforcement Learning. It aims to develop novel solutions for logic-enabled reasoning about RL-enabled ML systems. Concretely, the PhD research project is broadly organized into three main vectors :

1. First, to conduct an in-depth review of existing heuristic approaches for reasoning about RL, and to identify possible limitations of existing approaches.
2. Second, to develop a deep understanding of the work of the DeepLever team which has been working for several years on the other branches of ML including Neural Networks and statistical computational learning, on computing rigorous explanations.
3. Third, to develop formal tools for reasoning about Markov Decision Processes (MDPs), namely :
   (a) Generalize prime implicants of decision functions to the case of MDPs. One approach to investigate will be quantified functions representing strategies, similarly to what is common practice when solving quantified problems ;
   (b) Propose algorithms for computing logical formulations of MDPs behaviors ; and
   (c) Understand the practical limitations of computing compact logical formulations of MDPs, as well as the reasons behind their operation.

The first ideas are being tested on 2-person games and on examples of multi-agent path finding.

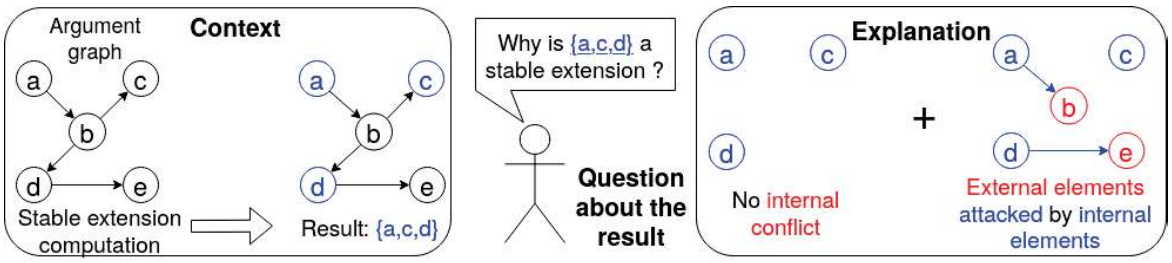

Figure 1.1 – An overview of computing explanations in abstract argumentation

Explainable AI for Intrusion Detection

Yannick Chevalier

Reflecting upon the usage of AI methods in the field of Intrusion Detection [23], Sommer and Paxson pointed out the gap between what can be offered by AI techniques for prediction and classification, and what is needed for effective intrusion detection. Among other, we shall name the need to build a system that distinguishes between anomalies and intrusions in a system, takes external descriptions of normal and intrusion behaviours into account and is able to explain its decisions to a human for further processing.

These considerations resonate with those expressed in [19] to define how a usable AI-based computer system should interact with a user, though with an emphasis on the system being an Advice Giver, to explain its decision, as much as an Advice Taker, to input external descriptions.

We built an intrusion detection system for simple networks in which the output of the learning is a set of first-order logic atoms that have to be satisfied by normal traffic [3]. This system is currently being expanded to prepare a background first-order logic theory that describes normal behaviours, and to construct abstract formulas describing the output of the learning phase.

Interacting a machine Learning system with an explicit reasoning system : Application on medical data

Mouna Sabrine Mayouf, Florence Dupin de Saint-Cyr

The PhD is about making interact a machine learning system with an explicit reasoning system for an application on medical data. This PhD started in December 2019.

A first research project has examined methodological aspects of the training procedure of neural networks in the context of a medical image classification problem. We have proposed a formalization of the data preparation. The formalism has allowed us to prove a number of useful properties of the training dataset used in the experiments, which in turn enhanced fairness of comparison and research transparency.

The second research question is concerning the conjecture that is, feeding a network with datasets of increasing magnification leverages high-level knowledge and helps the network to better classify. This hypothesis was confirmed by an experiment carried out on a dataset of breast cancer histopathological images. Results underline the importance of the order in which data is introduced to the neural network during the training phase. Extensive experiments done on the BreakHis dataset demonstrate that curriculum incremental learning reaches 98.76% accuracy for binary classification, while the best state-of-the-art approach only reaches 96.78%.

Concerning multi-class classification, curriculum incremental learning reaches 95.93% while the state-of-the-art approaches only reaches 95.49%. Also, both the computational time and the stabilization time of the learning process of the incremental curriculum learning approach are reduced (respectively by 6% and by more than 20%) as compared to a non curriculum learning approach.

We are currently working on a new way to use hierarchical constraints in order to guide the machine learning process. A first article has been accepted at the conference CAP'2021 [18] and a second article is under review for publication in an international journal [17].

Références

[1] A. Ignatiev Kuldeep S. Meel J. Marques-Silva M. Y. Vardi Aditya A. Shrotri, Nina Narodytska. Constraint-driven expla-

nations for black box ml models. In *Proc. of AAAI*, 2022.

[2] N. Asher, S. Paul, and Ch. Russell. Fair and Adequate Explanations. In *5th International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE 2021)*, volume 12844 of *LNCS*, Vienna (virtual), Austria, August 2021.

[3] Y. Chevalier. Data exchange for anomaly detection : The case of the can bus. In *Proceedings of the Conference on Artificial Intelligence for Defence*, 2021.

[4] M. C. Cooper and J. Marques-Silva. On the tractability of explaining decisions of classifiers. In *27th Int. Conf. on Principles and Practice of Constraint Programming, CP*, volume 210, 2021.

[5] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2) :321–357, 1995.

[6] X. Huang, Y. Izza, A. Ignatiev, and J. Marques-Silva. On efficiently explaining graph-based classifiers. In *Proc. of KR*, pages 356–367, 2021.

[7] A. Ignatiev, M. C. Cooper, M. Siala, E. Hebrard, and J. Marques-Silva. Towards formal fairness in machine learning. In *Proc. of the 26th Int. Conf. on Principles and Practice of Constraint Programming (CP)*, volume 12333 of *LNCS*. Springer, 2020.

[8] A. Ignatiev, Y. Izza, P. J. Stuckey, and J. Marques-Silva. Using maxSAT for efficient explanations of tree ensembles. In *Proc. of AAAI*, 2022.

[9] A. Ignatiev, N. Narodytska, N. Asher, and J. Marques Silva. From Contrastive to Abductive Explanations and Back Again.

[10] A. Ignatiev, N. Narodytska, and J. Marques-Silva. Abduction-based explanations for machine learning models. In *Proc. of AAAI*, 2019.

[11] A. Ignatiev, N. Narodytska, and J. Marques-Silva. On relating explanations and adversarial examples. In *Proc. of NeurIPS*, 2019.

[12] A. Ignatiev and J. P. Marques Silva. SAT-based rigorous explanations for decision lists. In *Proc. of the 24th Int. Conf. on Theory and Applications of Satisfiability Testing (SAT)*, volume 12831 of *LNCS*, pages 251–269, 2021.

[13] Y. Izza and J. Marques-Silva. On explaining random forests with SAT. In *Proc. of the Thirtieth Int. Joint Conference on Artificial Intelligence, IJCAI*, pages 2584–2591. ijcai.org, 2021.

[14] J. Marques-Silva, Th. Gerspacher, M. C. Cooper, A. Ignatiev, and N. Narodytska. Explaining naive bayes and other linear classifiers with polynomial time and delay. In *Proc. of NeurIPS*, 2020.

[15] J. Marques-Silva, Th. Gerspacher, M. C. Cooper, A. Ignatiev, and N. Narodytska. Explanations for monotonic classifiers. In *Proc. of the 38th Int. Conference on Machine Learning, ICML*, volume 139, pages 7469–7479, 2021.

[16] J. Marques-Silva and A. Ignatiev. Delivering trustworthy ai through formal XAI. In *Proc. of AAAI*, pages 3806–3814, 2022.

[17] M. Sabrine Mayouf and F. Dupin De Saint Cyr Bannay. Formalizing data preparation in curriculum incremental deep learning on breakhis dataset (revised version submitted to neurocomputing). 2021.

[18] M. Sabrine Mayouf and F. Dupin De Saint Cyr Bannay. Préparation efficace des données d'apprentissage. Application à la

classification d'images pour la détection du cancer du sein. In *Conférence sur l'Apprentissage Automatique (CAp 2021)*, Saint-Étienne (virtuel), France, 2021.

[19] J. McCarthy. Situations, actions, and causal laws. Technical Report TR AIM-002, Stanford University.

[20] N. Narodytska, A. A. Shrotri, K. S. Meel, A. Ignatiev, and J. Marques-Silva. Assessing heuristic machine learning explanations with model counting. In *Theory and Applications of Satisfiability Testing (SAT)*, volume 11628 of *LNCS*, pages 267–278. Springer, 2019.

[21] H. Prade and G. Richard. Explications analogiques. In *Workshop EXPLAIN'AI'22 @ EGC conf., Blois*, 2022.

[22] A. Rago, O. Cocarascu, C. Bechlivanidis, and F. Toni. Argumentation as a framework for interactive explanations for recommendations. In *Proc. of KR*, pages 805–815, 2020.

[23] R. Sommer and V. Paxson. Outside the closed world : On using machine learning for network intrusion detection. In *2010 IEEE Symposium on Security and Privacy*, 2010.

[24] A. Ignatiev M. Cooper N. Asher X. Huan, Y. Izza and J. Marques-Silva. Tractable explanations for d-DNNF classifiers. In *Proc. of AAAI*, 2022.