An examination of the effect of the inconsistency budget in weighted argumentation frameworks and their impact on the interpretation of deep neural networks*

Giulia Vilone[1,2,*,†], Luca Longo[1,2,†]

[1]Artificial Intelligence and Cognitive Load Research Lab
[2]School of Computer Science, Technological University Dublin

Abstract

Explaining the logic of a data-driven Machine Learning (ML) model can be seen as a defeasible reasoning process that is likely non-monotonic. This means a conclusion linked to a set of premises can be withdrawn when new information becomes available. Argumentation Theory (AT) formalises reasoning with a defeasible knowledge base. Abstract Argumentation Frameworks (AAF) organise conflicting arguments in a dialogical structure, allowing formal semantics to resolve conflicts. This study proposes an XAI method for automatically forming an AAF-based representation, using weighted attacks to model conflictual information. The concept of inconsistency budget is employed to eliminate the weakest attacks. Findings showed that the variation of the inconsistency budget could affect, albeit limited, the evaluation metrics computed over the resulting rulesets.

Keywords

Explainable artificial intelligence, Argumentation, Non-monotonic reasoning, Automatic attack extraction, Weighted argumentation frameworks, Inconsistency budget

1. Introduction

Numerous eXplainable AI (XAI) methods generate explanations of ML models in different formats (numerical, rules, textual, visual or mixed) [1, 2]. Rule-based explanations are considered naturally transparent and intelligible because they are a structured, compact, and intuitive format for reporting information [3]. Rule-based approaches usually consist of rulesets clarifying the relationships between the inputs of a model and its outputs. However, these approaches neither verify if rules are consistent with the background knowledge nor handle potential inconsistencies among rules [4, 5]. Understanding the inferential process of a model can be considered a non-monotonic reasoning process that allows the withdrawal of some conclusions,

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

*Corresponding author.
†These authors contributed equally.
giulia.vilone@tudublin.ie (G. Vilone); luca.longo@tudublin.ie (L. Longo)
https://giuliavilone.github.io/ (G. Vilone); lucalongo.eu/about (L. Longo)
0000-0002-4401-5664 (G. Vilone); 0000-0002-2718-5426 (L. Longo)

carried out by some rules, in light of new information [6]. Similarly, a model's predictions may be discarded when in conflict with the existing knowledge. This decision should follow a process grounded on logic and evidence. Argumentation studies how conflicting arguments, usually formalised with a first-order logical language, can be presented, supported or discarded in a defeasible reasoning process and investigates formal approaches to evaluate the validity of their conclusions [7]. AT provides the basis for implementing these processes computationally [8, 9], usually based on the notion of 'arguments' and 'attacks', often treated in an abstract way [10]. Some scholars proposed methods for automatically mining arguments and attacks from neural networks [11, 12]. However, more work needs to be done to assign weights to arguments or attacks extracted from a trained model in an automatic way [13]. This study focuses on automatically forming an argumentation framework consisting of rules and attacks among them, and it uses the concept of *inconsistency budget* [14] to determine a threshold for the strength of such attacks, and it investigates the impact of its variation to the resulting argumentation framework.

2. Related work

AT can generate effective explanations by translating a model's inferences in an argumentation process that shows, step by step, how it concludes sets of conflicting arguments [12, 15, 16, 17]. Formal non-monotonic logic studies formal frameworks to capture and represent defeasible inferences. A defeasible concept consists of a set of pieces of information, called *arguments*, that can be invalidated by adding new information [18]. Defeasible argumentation provides a sound formalisation for reasoning from a defeasible knowledge base [15]. This process frequently requires the recursive analysis of conflicting arguments in a dialectical setting to determine which arguments should be accepted or discarded [10]. Abstract Argumentation Theory (AAT) is the dominant paradigm, whereby arguments are abstractly considered in a dialogical structure. The AAT-based frameworks share a defeasible knowledge base made of arguments, a set of *attacks* to model conflicts between two arguments, and a formal *semantic* for conflict resolution that implements non-monotonicity in practice and assigns a dialectical status (accepted or rejected) to the arguments [16, 19]. [20, 14] assigned numeric weights to attacks, thus introducing Weighted Argumentation Frameworks (WAF). Weights must be positive, real values that assess the strength of the attack or, equivalently, measure the inconsistency between two arguments. The notion of *inconsistency budget* indicates how much inconsistency must be tolerated. Given an inconsistency budget β, all the attacks whose sum of weights is less than or equal to β can be disregarded. However, there are no indications of determining the value of the inconsistency budget to form a WAF with the optimal set of attacks and arguments.

3. Design

The experiment conducted as part of this research consists of training and explaining a model using a WAF. It contains six phases, as described below.

Phase 1: Dataset preparation. The first step was to choose a set of five training datasets containing multi-dimensional data handcrafted by domain experts and a categorical labelled

target variable. The experiment was conducted on the Adult, Avila, Bank, Credit Card Default and Letter Recognition public datasets downloaded from the UCI Machine Learning Repository[1].

Phase 2: Model training. A feed-forward neural network with two fully-connected hidden layers was trained on each dataset. The networks' hyper-parameters (optimiser, activation function, dropout rate, number of hidden neurons, and batch size) were tuned with a grid search to reach the highest prediction accuracy; the training process was early-stopped to prevent overfitting.

Phase 3: Automatic formation of a knowledge base. The ML models and the datasets were fed into a rule-extraction method, presented in [6] that generates a set of $IF - THEN$ rules using a two-step algorithm. Each rule corresponds to a 'defeasible' argument in the resulting WAF. The weighted attacks were automatically extracted from the generated rules by following the process proposed in [13]. Generally, attacks are binary relations between two conflicting arguments and can be of different kinds [8]. This study considers only the following two types: 1) *rebutting*, and 2) *undercutting attacks* [8].

Phase 4: Conflict evaluation. The weight of each attack measures the degree of inconsistency between pairs of arguments. This inconsistency is the difference in the number of instances supporting one of the two conflictual rules and belonging to their overlapping area [13]. The concept of 'inconsistency budget' was used to determine how much inconsistency must be tolerated [14]. In this study, the inconsistency budget varied between 10% and 90%.

Phase 5: Dialectical status and accrual of arguments. Given conflicting arguments, their acceptance status must be assigned. The *ranking-base categoriser* semantic [21, 22, 23] assigns a rank value to each argument by considering the number of its attacks. If there are multiple arguments with the highest rank, they are grouped into sets according to the conclusion they support [13]. The most credible set is the one with the highest cardinality. In the case of ties, no conclusion can be reached. This semantics does not consider the notion of weights of attacks.

Phase 6: Explainability Objective evaluation. Eight metrics were chosen to objectively and quantitatively measure the degree of explainability of the generated rule sets, per the evaluation approach presented in [13]. Objectivity is reached by excluding any human intervention in the evaluation process. *Number of rules* and *average rule length* assess the syntactic simplicity of the rules and must be minimised [24]. *Fraction of output classes* and *fraction of overlap* quantify the rules' clarity and coherence. The former should be as low as possible to avoid conflicts, whereas the latter must be maximised to guarantee that all the target classes are considered. A ruleset must also be *complete*, *correct*, *faithful* to the model's predictions, and *robust* to be a valid representation of a model's inferential process [25, 26, 27, 28, 24].

4. Results and conclusions

The variation in the value of the inconsistency budget has, generally speaking, a limited impact on the values of the metrics as the results remained almost unaltered throughout the five datasets (see Fig. 1). Completeness and fraction of classes remained constant at 100%, whereas the values of the other metrics for some datasets vary when the inconsistency budget goes above 50%. The inconsistency budget is the cause of these variations in the metrics as they correspond to a

[1]https://archive.ics.uci.edu/ml/index.php

sharp increase in the number of eliminated attacks (see Fig. 2). This supports the assumption that the number of attacks affects the robustness, correctness, fidelity and, sometimes, other metrics calculated over the predictions made by a WAF, even if this effect is limited in some instances. However, the limitations of this study in terms of the number and variety of models and datasets prevent reaching definitive conclusions. Further studies with datasets containing additional types of input data, such as texts and images, and ML models based on deeper neural networks or other learning architectures would tell if the few variations in the metric values are exceptions or if the inconsistency budget truly has an impact. Future work will extend this research study by using other semantics considering the attacks' weights.

Figure 1: Values of the objective metrics obtained over the three argumentation frameworks obtained by varying the inconsistency budget.

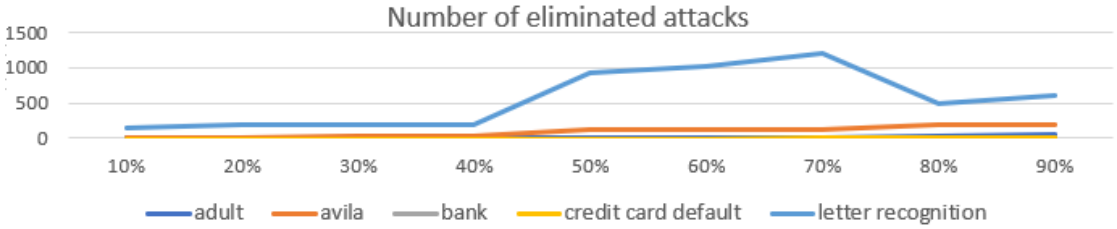

Figure 2: Number of eliminated attacks in the argumentation frameworks at the variation in the value of the inconsistency budget.

References

[1] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (2018) 93:1–93:42. doi:10.1145/3236009.

[2] G. Vilone, L. Longo, Classification of explainable artificial intelligence methods through their output formats, Machine Learning and Knowledge Extraction 3 (2021) 615–661.

[3] H. K. Dam, T. Tran, A. Ghose, Explainable software analytics, in: Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results, ACM, Gothenburg, Sweden, 2018, pp. 53–56.

[4] F. K. Došilović, M. Brčić, N. Hlupić, Explainable artificial intelligence: A survey, in: 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE, 2018, pp. 0210–0215.

[5] Z. C. Lipton, The mythos of model interpretability, Commun. ACM 61 (2018) 36–43.

[6] G. Vilone, L. Longo, A novel human-centred evaluation approach and an argument-based method for explainable artificial intelligence, in: IFIP International Conference on Artificial Intelligence Applications and Innovations, Springer, 2022, pp. 447–460.

[7] D. Bryant, P. Krause, A review of current defeasible reasoning implementations, The Knowledge Engineering Review 23 (2008) 227–260.

[8] L. Longo, Argumentation for knowledge representation, conflict resolution, defeasible inference and its integration with machine learning, in: Machine Learning for Health Informatics, Springer, 2016, pp. 183–208.

[9] L. Rizzo, L. Longo, An empirical evaluation of the inferential capacity of defeasible argumentation, non-monotonic fuzzy reasoning and expert systems, Expert Systems with Applications 147 (2020) 113220.

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

[11] O. Cocarascu, K. Cyras, F. Toni, Explanatory predictions with artificial neural networks and argumentation, in: Proceedings of the 2nd Workshop on Explainable Artificial Intelligence (XAI 2018), 2018.

[12] O. Cocarascu, F. Toni, Argumentation for machine learning: A survey., in: COMMA, 2016, pp. 219–230.

[13] G. Vilone, L. Longo, A global model-agnostic xai method for the automatic formation of an abstract argumentation framework and its objective evaluation, in: 1st International Workshop on Argumentation for eXplainable AI co-located with 9th International Conference on Computational Models of Argument (COMMA 2022), CEUR Workshop Proceedings, 2022, p. 2119.

[14] P. E. Dunne, A. Hunter, P. McBurney, S. Parsons, M. Wooldridge, Weighted argument systems: Basic definitions, algorithms, and complexity results, Artificial Intelligence 175 (2011) 457–486.

[15] S. A. Gómez, C. I. Chesnevar, Integrating defeasible argumentation and machine learning techniques: A preliminary report, in: In Procs. V Workshop of Researchers in Comp. Science, 2003, pp. 320–324.

[16] S. Modgil, F. Toni, F. Bex, I. Bratko, C. I. Chesnevar, W. Dvořák, M. A. Falappa, X. Fan, S. A. Gaggl, A. J. García, et al., The added value of argumentation, in: Agreement technologies, Springer, 2013, pp. 357–403.
[17] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and explainable artificial intelligence: a survey, The Knowledge Engineering Review 36 (2021) e5.
[18] L. Longo, L. Rizzo, P. Dondio, Examining the modelling capabilities of defeasible argumentation and non-monotonic fuzzy reasoning, Knowledge-Based Systems 211 (2021) 106514.
[19] S. A. Gómez, C. I. Chesnevar, Integrating defeasible argumentation with fuzzy art neural networks for pattern classification, Journal of Computer Science & Technology 4 (2004) 45–51.
[20] P. E. Dunne, A. Hunter, P. McBurney, S. Parsons, M. J. Wooldridge, Inconsistency tolerance in weighted argument systems., in: AAMAS (2), 2009, pp. 851–858.
[21] L. Amgoud, J. Ben-Naim, Ranking-based semantics for argumentation frameworks, in: International Conference on Scalable Uncertainty Management, Springer, 2013, pp. 134–147.
[22] L. Amgoud, J. Ben-Naim, D. Doder, S. Vesic, Ranking arguments with compensation-based semantics, in: Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning, 2016, pp. 12–212.
[23] P. Besnard, A. Hunter, A logic-based theory of deductive arguments, Artificial Intelligence 128 (2001) 203–235.
[24] H. Lakkaraju, S. H. Bach, J. Leskovec, Interpretable decision sets: A joint framework for description and prediction, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, San Francisco, California, USA, 2016, pp. 1675–1684.
[25] G. Bologna, Y. Hayashi, A comparison study on rule extraction from neural network ensembles, boosted shallow trees, and svms, Applied Computational Intelligence and Soft Computing 2018 (2018). doi:10.1155/2018/4084850.
[26] C. Ferri, J. Hernández-Orallo, M. J. Ramírez-Quintana, From ensemble methods to comprehensible models, in: International Conference on Discovery Science, Springer, Lübeck, Germany, 2002, pp. 165–177. doi:10.1007/3-540-36182-0_16.
[27] A. A. Freitas, Are we really discovering interesting knowledge from data, Expert Update (the BCS-SGAI magazine) 9 (2006) 41–47.
[28] A. Ignatiev, Towards trustable explainable AI, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI, Yokohama, Japan, 2020, pp. 5154–5158. doi:10.24963/ijcai.2020/726.