Computational Models of Argument
F. Toni et al. (Eds.)

doi:10.3233/FAIA220166

An Argumentative Explanation of Machine Learning Outcomes[1]

Stefano BISTARELLI [a,2], Alessio MANCINELLI [a], Francesco SANTINI [a,2], and Carlo TATICCHI [a,2,3]
[a] *Dipartimento di Matematica e Informatica, Università degli Studi di Perugia, Italy*

Keywords. Computational argumentation, explainability, machine learning.

The black box model used in Machine Learning is considered one of the major problems in the application of Artificial Intelligence techniques [1] as it makes machine decisions non-transparent and often incomprehensible even to experts or developers themselves. In this paper, we provide an argumentative interpretation of both the training process and the results predicted. The goal is to build a Bipolar Argumentation Framework (BAF) [2] showing the dialectical reasoning behind the assignment of a certain class to a given record. Since we make assumptions neither on the dataset nor on the algorithm used, the presented procedure can be applied to existing models without the need for further adjustments. To illustrate our proposal, we use the *Titanic* dataset from www.kaggle.com, which contains records relating to people involved in the Titanic disaster. We consider three categorical features, namely *Survived* (the class to predict, with value 1 if the person survived or 0, otherwise), *Pclass* (ticket class among 1, 2 and 3) and *sex* (0 for woman and 1 for man), and two numerical features: *Age* (passenger age, ranging from 0.17 to 76) and *Fare* (passenger fare with values from 0 to 512). In the following, we describe the step our procedure goes through in order to find an explanation for the class *Survived=1*.

Dataset Clustering. In the first step, starting from the input dataset, we create a new clustered dataset in which numerical features are split into categories that group ranges of values to obtain a more appropriate and concise explanation.

BAF Generation. Then we build a BAF based on the correlation matrix computed among the features. By construction, the obtained BAF only has symmetric relations.

Breaking Complete Symmetry. Given the correlation matrix, we apply a procedure that removes symmetric edges from the BAF to establish a causal relationship between features. In particular, we use the conditional probability [3] computed for arguments which attack/support each other. We choose the minimum values possible that keep the graph connected.

Computing Extensions. To identify the set of arguments which are more likely to be accepted, we compute the semi-stable extensions [4] of the previously obtained

[1] This work has been partially supported by: GNCS-INdAM, CUP E55F22000270001; Project RACRA - funded by Ricerca di Base 2018-2019, Univeristy of Perugia; Project BLOCKCHAIN4FOODCHAIN: funded by Ricerca di Base 2020, Univeristy of Perugia; Project DopUP - REGIONE UMBRIA PSR 2014-2020.

[2] The author is a member of the INdAM Research group GNCS and of Consorzio CINI.

[3] Corresponding Author: Carlo Taticchi, Università degli Studi di Perugia; E-mail: carlo.taticchi@unipg.it.

framework and then we use the tool described in [5] to find, for each of them, its probability of being admissible. In our example, we obtain the following extension, which is semi-stable and also admissible with probability 1 (the highest possible).

$$Age<0.96,\ Fare\geq 10.48,\ Pclass=1,\ Sex=0,\ \boldsymbol{Survived=1}$$

Building the Explanation Tree. Finally, starting from the arguments of the selected extension, we produce the explanation tree of Figure 1, where accepted arguments are highlighted in green and rejected ones in red.

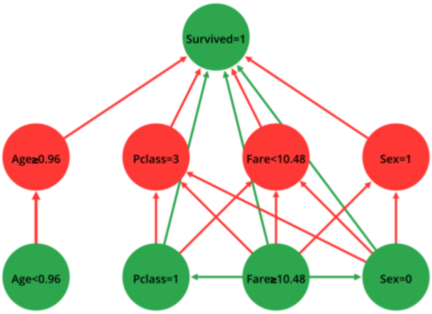

Figure 1. An explanation tree for the class *Survived=1* of the Titanic dataset.

Looking at the obtained explanation we can conclude, for instance, that the person in question survived because "she is a woman (*Sex=0*), with a paid ticket (*Fare≥10.48*) and travelling first class (*Pclass=1*)". Indeed, arguments representing those features in Figure 1 attack other arguments that are against the assignment of the class *Survived=1*, standing in turn for being male (*Sex=1*) and having a third-class ticket (*Pclass=3*) with a low fare (*Fare<10.48*).

In future work, alternative techniques could be applied to break the symmetry of the graph to obtain a causal relationship between arguments. Furthermore, particular attention could be paid to simplifying the explanation provided, including notions of symmetry and interchangeability between arguments, as well as applying Natural Language Processing to provide a further textual explanation.

References

[1] von Eschenbach WJ. Transparency and the Black Box Problem: Why We Do Not Trust AI. Philosophy & Technology. 2021 Dec;34(4):1607-22.

[2] Amgoud L, Cayrol C, Lagasquie-Schiex M, Livet P. On bipolarity in argumentation frameworks. Int J Intell Syst. 2008;23(10):1062-93.

[3] Casella G, Berger RL. Statistical inference. vol. 2. Duxbury Pacific Grove, CA; 2002.

[4] Baroni P, Caminada M, Giacomin M. An introduction to argumentation semantics. Knowl Eng Rev. 2011;26(4):365-410.

[5] Bistarelli S, Mantadelis T, Santini F, Taticchi C. Using MetaProbLog and ConArg to compute Probabilistic Argumentation Frameworks. In: Proceedings of the 2nd Workshop on Advances In Argumentation In Artificial Intelligence. vol. 2296 of CEUR Workshop Proceedings. CEUR-WS.org; 2018. p. 6-10.