A Novel Model-Agnostic xAI Method Guided by Cost-Sensitive Tree Models and Argumentative Decision Graphs

Marija Kopanja[1,2,*]

[1]*Faculty of Sciences, University of Novi Sad, Trg Dositeja Obradovića 3, 21000 Novi Sad, Serbia*
[2]*BioSense Institute, Dr Zorana Djindjića 1, 21000 Novi Sad, Serbia*

Abstract

In recent years there is increasing demand for comprehension and explainability of the inferences machine learning (ML) models make. Many explainable artificial intelligence (xAI) methods have been introduced as a tool for better understanding of inference process of complex AI models. The doctoral research aims to develop a new model-agnostic xAI framework for classification tasks by using cost-sensitive decision trees and argumentative decision graphs. From the classification problem point of view, especially if a dataset is imbalanced, the cost-sensitive decision tree (CSDT) method can be used for generating an acceptably accurate ML model by taking imbalance ratio into the consideration during the tree building procedure. On the other side, from the explainability perspective, generated cost-sensitive tree model can be more comprehensible compared to the tree model generated using traditional (cost-insensitive) decision tree learning algorithm, due to smaller tree size of the cost-sensitive tree. However, to have more plausibly accurate ML model for given imbalanced classification task, deep learning algorithms could be applied, leading to more complex, non-linear models whose decision-making process is hard to understand and explain. For such complex models, we can create surrogate model that will approximate the predictions of the underlying model as accurately as possible, while at the same time being interpretable and easy to explain. For the purpose of creating surrogate model, a cost-sensitive decision tree learning algorithm can be used. By having a CSDT model, it is possible to obtain explanation for any sample as a rule extracted from the tree. Thereby, we can consider cost-sensitive tree as a rule-extraction xAI method. Current research show that argumentation graph can represent the logic of the complex model with fewer rules than a decision tree. The aim of the study is to investigate possible ways of transforming cost-sensitive tree model into an argumentative decision graph in order to create a more concise structure that should be more understandable. The final step of generating argument-based explanations is evaluation by using both quantitative and human-center analysis.

Keywords

Explainable Artificial Intelligence, Model Agnostic Explanations, Explainable Surrogate Models, Cost-sensitive Decision Tree, Argumentation, Machine Learning

1. Introduction and research motivation

Predictive machine learning (ML) models play a crucial role in various fields, from finance, agriculture, to healthcare. As the availability of data exponentially increases, ML methods,

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

[*]Corresponding author.
marija.kopanja@biosense.rs (M. Kopanja)
0000-0003-1176-5677 (M. Kopanja)

particularly deep learning methods, have led to the creation of powerful models. However, many of these models are characterized by complex, non-linear structures that can be challenging to interpret and explain. One of the important factors when using the ML model in production regardless of the domain of application, or in research, is the interpretability of the model [1]. Many explainable artificial intelligence (xAI) methods have been introduced as a tool for a better understanding of the inference process of complex ML models. There is a plethora of xAI methods and there have been many attempts to make a unified division of xAI methods [2, 3, 4]. Some approaches in the categorization of the xAI methods focus on the type of input data used to train the ML model, others focus on the internal mechanisms of the xAI method, while some focus on the scope i.e. whether the xAI method generates local and/or global explanations. Another way to segregate the xAI method is by determining whether the method is post-hoc or ante-hoc. The former group of xAI methods enable an understanding of the black-box model a posteriori, while the latter group tries to make the ML model naturally explainable. The advantage of any post-hoc method is that there is no influence on the performance of the black-box model which is important due to a trade-off between predictive performance and transparency, as the two objectives are conflicted [5]. This problem any many other challenges related to xAI are discussed in several papers [6, 7, 8].

The doctoral research aims to develop a new post-hoc, model-agnostic xAI framework for classification tasks by using cost-sensitive decision tree (CSDT) method and argumentative decision graphs. Creating the new method is motivated by the fact that generating a posteriori explanation can be the only solution for explaining already trained black-box ML models. The method to be developed will be model-agnostic, hence without requirements in terms of understanding the inner workings of the ML model to be explained. For any complex model, it can be created a surrogate model that will approximate the predictions of the underlying model as accurately as possible, while at the same time being interpretable and easy to explain. To create a surrogate model, a CSDT learning algorithm can be used. By having a CSDT model, it is possible to obtain an explanation for any sample as a rule extracted from the tree. Thereby, we can consider a CSDT as a rule-extraction xAI method. Although tree-based models are considered as naturally transparent and interpretable [6], for a layman it can be difficult to comprehend explanations given by a tree model, especially if the tree is large. An extracted set of rules from the tree model should contain as few concise and short rules for as many samples as possible [9]. Current research [10] shows that an argumentation graph can represent the logic of a complex model with fewer rules than a decision tree. In our framework one of the objectives is to use CSDT model since generated CSDT model can be more comprehensible compared to the tree model generated using a traditional (cost-insensitive) decision tree learning algorithm [11], due to the smaller tree size of the cost-sensitive tree. The extracted rules from any tree-based model should mimic the inferential process of a complex ML model [5, 12, 9]. To bridge the gap between lack of transparency and non-linearity of complex ML model, the aim of the research is to develop new xAI method that will be based on rules extracted from a surrogate CSDT model, further transforming the rules into an argumentative decision graph.

2. Key related works that frame the research

2.1. Surrogate xAI models

One of the most popular model-agnostic xAI approaches is creating surrogate model for the complex ML model to be explained [3]. The surrogate model is created to accurately approximate the predictions of the complex, black-box ML model, while still being interpretable. The only requirement for the approach is to have training data and the predictions of the model to be explained. The surrogate model can be global or local, depending if the original dataset is used for training the model or just a subset of the original data. For example, the LIME method [13] is a local post-hoc model-agnostic explanation method, meaning it generates an explanation by using a new set of samples in the proximity of the sample to be explained and training a local interpretable linear model. There are many studies that tried to improve the LIME and resolve its issues with stability (problem of generating the same explanations for the same sample in several runs) and local fidelity (problem when learned explanation model is not a good local approximation of the model being explained), such as the ALIME method ([14]) that uses autoencoders for assigning weights for samples and uses linear model as a local surrogate model. Explanations provided by local interpretable model in view of the feature scores and prediction probability can be hard to understand and interpret since the feature scores do not add up to the prediction probability. Therefore, other interpretable models such as tree-based models could be used. In [15] is proposed new approach tree-ALIME, modified version of ALIME, which uses a decision tree as an interpretable model. As their results of evaluating tree-ALIME show, using a decision tree model as a local interpretable model is promising. However, the results show that using a decision tree model instead of a linear model, did not improve local fidelity probably due to a simple decision tree model (maximal depth of the tree is set to be 5) and a tendency of tree models to overfit the data. More importantly, regarding interpretability, the decision tree model gained significantly better results compared to the linear model. Therefore, other tree-based algorithms can be used in the proposed approach to tackle all aforementioned challenges. In the abundance of tree-based models, it is possible to use the CSDT in tree-ALIME approach as the local interpretable model. On the other side, any decision tree algorithm including the CSDT algorithm, can be used to create a global surrogate model which might be an approach more aligned with the aim of this research.

2.2. Cost-sensitive decision tree

The cost-sensitive decision tree (CSDT) method [16] is a ML algorithm for generating a tree model by considering the cost matrix during the tree-building procedure. The CSDT method belongs to the group of cost-sensitive learning methods ([17]), that can be used in the more narrow, imbalanced learning framework. This approach can be seen as an algorithm-level solution for the class imbalance problem, since there is an adaptation of existing classification learning algorithm to improve performance with regards to the minority class. On the other hand, a data-level solutions assume different rebalancing techniques to make data distribution more balanced, having its limits and costs. Therefore, using algorithm-level solutions such as the CSDT model might be more convenient option from the classification problem point of view. On the other hand, from the explainability perspective, a cost-sensitive tree model can be more comprehensible compared to the traditional decision tree learning algorithm.

The tree structure of the model enables us to create explanation for each sample by following the path from the root node to the leaf node of the tree. To create a CSDT model it must be given the test set, the prediction labels of the corresponding test set obtained from the black box ML model to be explained, and the cost matrix. In general, a cost matrix can be either class-dependent (all samples from the same class have the same cost matrix) or sample-dependent (each sample has its cost matrix). Having proper cost-matrix defined is essential for cost-sensitive tree-building process, since the CSDT algorithm chooses a feature that reduces the misclassification cost the most. That is, the CSDT uses the cost-sensitive splitting criterion and unlike traditional decision tree, a cost-sensitive tree will classify the sample in the region to the least costly class. The resulting product is the tree object, as in any other tree-based ML algorithm, that is considered naturally transparent and explainable. Nevertheless, any tree model can be hard to understand if the model is deep, and this might be the case if the cost-sensitive tree-model is used as a surrogate model. To be reliable, a CSDT surrogate model must achieve high performance and be able to predict the same output as the complex ML model before providing explanations. Therefore, the generated tree model might be deep and hence it can be hard to comprehend its inference process. All things considered, the doctoral research broaden the scope into the argumentation framework since rules can be seen as arguments in the filed of argumentation [9].

2.3. Argumentation framework

Argumentation is a multidisciplinary subfield of AI that studies how arguments can be presented in a defeasible reasoning (a formalism for non-monotonic reasoning) process and how to evaluate the validity of the conclusions reached at the end of the reasoning process [10, 18, 19, 20]. Argument-based systems are typically build upon multi-layer schema [21, 18, 19]. Argumentation has several important concepts: arguments, attacks and semantics [21]. The arguments are rules and attacks are binary relations between two conflicting rules (arguments) and three classes of conflicts can be distinguished [21]. A fundamental feature of argument-based system is ability to determine the success of an attack [10]. For example to decide if an attack is valid the strengths of arguments or attacks can be used [21].

Argumentative decision graphs (ADGs) have a rule-based structure where each argument has a single premise and a conclusion. The well-formed ADG can be extracted from a decision tree, by taking each terminal node in the tree to generate a predictive argument in the ADG, while non terminal nodes could be used as non predictive arguments [22]. The attacks could be generated between arguments with different features and conclusions that are in disjoint paths and lead to distinct terminal nodes.

In the [22] is proposed new argumentative decision graph method, xADG (extend argumentative decision graph), where the emphasis was on decision trees and argumentative models. They showed that based on tree model, proposed method could create extended argumentative decision graph of equivalent inferential capability that could be perceived as more understandable. It is important that derived argumentative model is guaranteed to maintain the same inferential capability, still being smaller in terms of a size. They analysed whether reasonably smaller structures, in terms of number of arguments/attacks and amount of argument supports,

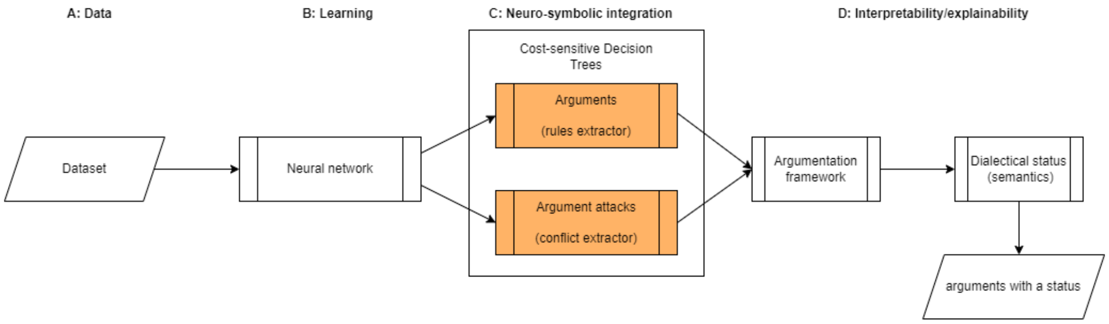

Figure 1: Conceptual framework of the new post-hoc, model-agnostic xAI method.

can be achieved for classification tasks. Their results suggest that leveraging the structure and inferential capability of tree model with proposed novel framework for structured argumentation could be good alternative for automating the creation of reasonably sized argumentation framework.

3. Specific research questions, hypothesis and objectives

The doctoral research will be carried out in a several phases described in the following paragraphs and depicted in the diagram (Figure 1).

Having the dataset spitted onto the train and test subset, the black-box ML model is trained on the train subset and evaluated on the test. Next step is to provide understanding into the inference process of the the complex ML model which will be done in phases. First phase is creating a surrogate model by using inherently interpretable cost-sensitive decision tree model.

The second phase has an objective to transform obtained rules from CSDT into argument-based representation. The process can be broken down into five layers [21, 18, 19]: 1. definition of the internal structure of arguments 2. definition of conflicts between arguments 3. evaluation of conflicts and definition of valid attacks 4. definition of the dialectical status of arguments 5. accrual of acceptable arguments. One of our research questions related to argumentation and described multi-layer schema is: whether the weighted notion of argument or attack should be considered in our framework, where weights would represent the strength of the argument or attack measured by considering misclassification cost reduction? For example, if two paths (rules) in the tree model are conflicted, weights could be computed as the misclassification cost of samples belonging to the intersection of the covers of two conflicting rules that are assigned by the model to the same target class of the conclusion of the attacking rule.

Given a set of arguments with defined attacks, a further decision that must be made is which arguments can be accepted. An algorithm designed to produce a set of acceptable and conflict-free arguments is called semantics [18]. Different semantics, such as grounded or preferred, can be used, leading to a set of arguments with a status (rank). In [22] is shown that rules from a tree model exploited by an extension based semantics, such as grounded, results in ADG with the same set of inferences as the tree. Therefore, another question is related to the choice of semantics designed for handling the (weighted) argumentation framework.

The extend argumentative decision graph (xADG), proposed in [22] is as new framework that allows for arguments to use boolean logic operators and multiple premises within their internal structure, resulting in more concise argumentative graphs that may be easier for users to understand. The xADG of equivalent inferential capability as ADG, is formed by performing a set of modifications. We aim to test if the proposed framework xADG can be applied to ADG built from CSDT and what modifications are needed if weighted argumentation is going to be used. Therefore another research question we aim to answer is whether using CSDT instead of the decision tree algorithm to derive an argumentative decision graph, would results in the more comprehensible graph.

4. Current results and next steps

To date, the CSDT models are trained on various datasets with different class imbalance ratios. The current results show that cost-sensitive tree model is less complex compared to the traditional decision-tree model, for the same tree depth, without implementation of a pruning procedure [11]. In further work, we aim to extend the number of datasets used for the comparison purposes in order to test if rules extracted from a cost-sensitive tree model are consistently shorter that the rules extracted from a traditional decision tree model.

In the next step, a CSDT will be created as surrogate model for some complex ML model such as deep neural network model. Afterwards, the CSDT model should be transformed into argumentative decision graph to generate simpler rules that are potentially more comprehensible as is done in [22].

The final step of generating argument-based explanations will be evaluation. In general, two ways of evaluating interpretability of the model can be distinguished: quantitative and human-centered evaluations. The latter can include domain experts and/or people unfamiliar with concepts such as ML and xAI, in order to evaluate obtained explanations provided to individuals with diverse knowledge. As is done in the studies [9, 10], we can select several metrics to quantitatively assess the degree of explainability of the rules extracted from the CSDT and the rules of argumentation-based graph. For human-centred evaluation purposes of explanations produced, in future work the human-centred psychometric test [23] could be used. Developed argument-based model-agnostic xAI method should also be compared to other rule-based and argument-based xAI methods [10, 22].

5. Final contribution

The end product of the doctoral research is post-hoc model-agnostic argument-based xAI method developed by extraction of rules and their conflicts from CSDT models and their integration into an argumentation framework that can serve as a mechanism for interpreting and explaining the inferential process of complex ML models. Leveraging the structure and inferential capability of CSDT with argumentation decision graph could be promising direction in automating the creation of argumentation framework with reasonable size that will be more easy to comprehend by the end-users.

Acknowledgments

Supported by ANTARES project that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement SGA-CSA No. 739570 under FPA No. 664387, https://doi.org/10.3030/739570 and Ministry of Education, Science and Technological Development of the Republic of Serbia, grant agreement 451-03-47/2023-01/200358

References

[1] C. Molnar, G. Casalicchio, B. Bischl, Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges, Springer International Publishing, 2020, p. 417–431. doi:10.1007/978-3-030-65965-3_28.

[2] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2018) 42. doi:10.1145/3236009.

[3] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (xai), IEEE Access PP (2018) 1–1. doi:10.1109/ACCESS.2018.2870052.

[4] G. Vilone, L. Longo, Classification of explainable artificial intelligence methods through their output formats, Machine Learning and Knowledge Extraction 3 (2021) 615–661. doi:10.3390/make3030032.

[5] F. K. Došilović, M. Brčić, N. Hlupić, Explainable artificial intelligence: A survey, in: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 0210–0215. doi:10.23919/MIPRO.2018.8400040.

[6] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, W. Samek, Explainable AI Methods - A Brief Overview, Springer International Publishing, Cham, 2022, pp. 13–38. doi:10.1007/978-3-031-04083-2_2.

[7] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. D. Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, S. Stumpf, Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions, Information Fusion 106 (2024) 102301. doi:https://doi.org/10.1016/j.inffus.2024.102301.

[8] L. Longo, R. Goebel, F. Lecue, P. Kieseberg, A. Holzinger, Explainable artificial intelligence: Concepts, applications, research challenges and visions, in: A. Holzinger, P. Kieseberg, A. M. Tjoa, E. Weippl (Eds.), Machine Learning and Knowledge Extraction, Springer International Publishing, Cham, 2020, pp. 1–16. doi:10.1007/978-3-030-57321-8_1.

[9] G. Vilone, L. Longo, A quantitative evaluation of global, rule-based explanations of post-hoc, model agnostic methods, Frontiers in Artificial Intelligence 4 (2021) 160. doi:10.3389/frai.2021.717899.

[10] G. Vilone, L. Longo, A global model-agnostic xai method for the automatic formation of an abstract argumentation framework and its objective evaluation, volume 3209, CEUR-WS, 2022. doi:https://doi.org/10.1007/978-3-031-04083-2_2, publisher Copyright: © 2022 Copyright for this paper by its authors.; 1st International Workshop on Argumentation for eXplainable AI, ArgXAI 2022 ; Conference date: 12-09-2022.

[11] M. Kopanja, S. Hačko, S. Brdar, M. Savić, Cost-sensitive tree shap for explaining cost-sensitive tree-based models, Computational Intelligence 40 (2024) e12651. doi:https://doi.org/10.1111/coin.12651.

[12] E. Mekonnen, P. Dondio, L. Longo, Explaining deep learning time series classification models using a decision tree-based post-hoc xai method, volume 3554, CEUR-WS, 2023. doi:https://doi.org/10.21427/9YKT-WZ47, publisher Copyright: © 2023 CEUR-WS. All rights reserved.;

Joint 1st World Conference on eXplainable Artificial Intelligence: Late-Breaking Work, Demos and Doctoral Consortium, xAI-2023: LB-D-DC ; Conference date: 26-07-2023 Through 28-07-2023.

[13] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 1135–1144. doi:10.1145/2939672.2939778.

[14] S. M. Shankaranarayana, D. Runje, Alime: Autoencoder based approach for local interpretability, in: H. Yin, D. Camacho, P. Tino, A. J. Tallón-Ballesteros, R. Menezes, R. Allmendinger (Eds.), Intelligent Data Engineering and Automated Learning – IDEAL 2019, Springer International Publishing, Cham, 2019, pp. 454–463. URL: https://api.semanticscholar.org/CorpusID:202539758.

[15] N. Ranjbar, R. Safabakhsh, Using decision tree as local interpretable model in autoencoder-based lime, 2022. URL: https://arxiv.org/abs/2204.03321. arXiv:2204.03321.

[16] B. A. Correa, Example-dependent cost-sensitive decision trees, Expert Systems with Applications 42 (2015) 6609–6619. doi:10.1016/j.eswa.2015.04.042.

[17] C. P. Elkan, The foundations of cost-sensitive learning, in: International Joint Conference on Artificial Intelligence, 2001. URL: https://api.semanticscholar.org/CorpusID:16149383.

[18] L. Rizzo, L. Longo, A qualitative investigation of the explainability of defeasible argumentation and non-monotonic fuzzy reasoning, in: Proceedings for the 26th AIAI Irish Conference on Artificial Intelligence and Cognitive Science Trinity College Dublin, Dublin, Ireland, December 6-7th, 2018., 2018, pp. 138–149. doi:https://doi.org/10.21427/tby8-8z04.

[19] L. Longo, L. Rizzo, P. Dondio, Examining the modelling capabilities of defeasible argumentation and non-monotonic fuzzy reasoning, Knowledge-Based Systems 211 (2021) 106514. doi:https://doi.org/10.1016/j.knosys.2020.106514.

[20] 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 (2020) 113–220. doi:https://doi.org/10.1016/j.eswa.2020.113220.

[21] L. Longo, Argumentation for Knowledge Representation, Conflict Resolution, Defeasible Inference and Its Integration with Machine Learning, volume 9605, 2016, pp. 183–208. doi:10.1007/978-3-319-50478-0_9.

[22] L. Rizzo, A Novel Structured Argumentation Framework for Improved Explainability of Classification Tasks, Springer Nature Switzerland, 2023, p. 399–414. doi:10.1007/978-3-031-44070-0_20.

[23] G. Vilone, L. Longo, Development of a Human-Centred Psychometric Test for the Evaluation of Explanations Produced by XAI Methods, 2023, pp. 205–232. doi:10.1007/978-3-031-44070-0_11.