Semantic Explanations and Reasoning for Machine Learning Models: An Ontology-Based Approach

Laura Papi[1]

[1]Sapienza University of Rome, Italy

Abstract

In a context where Machine Learning models are becoming increasingly widespread and complex, the importance of techniques to make these models less *opaque* is widely recognized. The research fields of Neuro-symbolic Artificial Intelligence and Explainable Artificial Intelligence both revolve around techniques to increase the transparency of complex Machine Learning systems. The research proposed in this paper aims at defining a novel framework to exploit symbolic structures, in particular ontologies, to provide innovative solutions to the problem of understanding Machine Learning models. The core of our proposed approaches is represented by the Ontology Based Data Management paradigm, which is exploited to define a relation between the sub-symbolic data samples used in Machine Learning models, and the symbolic knowledge possessed by domain experts, formalized in an ontology. The key innovation of the presented framework is the definition of a novel form of *mapping*, which allows to reconcile symbolic and sub-symbolic knowledge. In particular, leveraging this framework, we are interested in studying two problems: *formulating explanations* for the behavior of models, and *performing reasoning* over these Machine Learning systems. We argue that investigating these problems could lead to valuable contributions to the research field.

Keywords

Neuro-symbolic AI, Ontology Based Data Management, Knowledge Representation, Explainable AI

1. Introduction

In recent years, the use of Machine Learning (ML) models spread across numerous domains, many of which involve significant ethical implications. As a result, the problem of *opacity* and low interpretability of these models has become a crucial area of interest. In the Artificial Intelligence (AI) research field, two of the branches that investigate this issue are Explainable AI and Neuro-symbolic AI. Explainable AI focuses on techniques to provide explanations to the decisions of ML models, usually classifiers. Neuro-symbolic AI instead, is the research field that studies how to integrate symbolic domain knowledge with sub-symbolic data representations used by ML models, as neural networks. The work presented in this paper falls in both these research fields, as we aim to exploit the Neuro-symbolic AI approach towards model interpretability and to provide symbolic explanations to the behavior of the models.

Specifically, the key idea of our proposed approach lies in the use of ontologies to represent the domain knowledge, and in employing the Ontology Based Data Management (OBDM) paradigm to reconcile the sub-symbolic representation of data sample with the symbolic domain knowledge of the ontology. The main advantage of this novel approach is represented by the high level languages that can be used to formalize the logical rules. Ontologies, by definition, formalize the knowledge in terms of concept and roles, which are particularly close to the human approach to reasoning. We argue that this approach could contribute to the field of Neuro-symbolic AI, making the symbolic components of these systems, such as logical rules and explanations, easier to define, manipulate, and understand, therefore enhancing interpretability, transparency and trustworthiness of the models.

The first objective we aim to achieve is the definition of a novel formal framework, based on the OBDM paradigm, that allows to reconcile the numerical representation of the data samples with the symbolic knowledge defined by the ontology. Then, we aim to study two main applications of this framework: *performing reasoning* over ML models and *providing semantic explanations* for their behavior.

SEBD 2025: 33rd Symposium on Advanced Database System, June 16-19, 2025 - Ischia, Italy

papi@diag.uniroma1.it (L. Papi)
0009-0003-2281-9500 (L. Papi)

The rest of the paper is organized as follows. In Section 2 we review the main techniques that form the foundation of this research. Section 3, which constitutes the core of this paper, is structured in three paragraphs: first outlines the open questions we aim to address; then presents the framework we intend to define; lastly discusses two interesting applications for the framework, along with some preliminary results, and future works. Finally, Section 4, concludes the paper by providing an overview of the presented research, its main motivations and objectives.

2. Background and Related Works

As mentioned in Section 1, the research presented in this paper falls under both the Neuro-symbolic AI and Explainable AI categories, and exploits the OBDM paradigm to address the described problems. In the following we provide a brief overview of these fields, together with some of their most recent and prominent approaches.

Neuro-symbolic AI: integrating symbolic knowledge with ML models
In the early days of AI research, symbolic approaches represented the dominant paradigm. However, in recent years, the most popular approaches to AI are sub-symbolic, i.e., focused on learning implicit data representations rather than using formal symbols and logical rules. Bridging the gap between these two approaches is at the core of Neuro-symbolic AI, which aims to obtain systems that leverage the strengths of both paradigms [1]. In particular, sub-symbolic systems, as neural networks, are able to solve highly complex problems by learning implicit patterns in the data, that would be much more difficult to explicitly formalize using semantic rules. Symbolic techniques, on the other hand, use human-readable representations of knowledge and data, making these models much more interpretable. Combining symbolic and sub-symbolic approaches offers several promising strategies, which typically include neural networks as the sub-symbolic component, and differ by the way symbolic strategies are integrated within the network [2]. Among the various contexts where Neuro-symbolic AI shows potential, our primary focus is on its ability to make ML models more understandable and interpretable.

Explainable AI: computing explanations for the behavior of ML models
This field of research emerged in response to the growing complexity of ML models, which resulted in highly *opaque* models, often treated as *black-boxes*. However, in recent years, these models became widespread, even in high-risk contexts, and their opacity raised significant ethical concerns [3, 4, 5]. This has led to a growing interest in techniques that aim to explain the behavior of these models, or offer insights into their internal processes [6].

The majority of the studies on explanations assume the model to explain is a classifier, and the subject of the explanation is its output decision. The most notable techniques are based on the following definitions: *local explanations* provide explanations for individual predictions, i.e., the output computed for a specific input instance; *global explanations*, instead provide explanations for whole classes, hence focusing on the behavior of the classifier in general [7]; *abductive explanations* [8], also known as *prime implicants* [9] or *sufficient reasons* [10], focus on answering the question of why an instance is classified in a specific way; *contrastive*, or *counterfactual*, explanations, on the other hand, focus on the question of why an instance isn't classified differently [11]; *post-hoc explanations* consider the model as a black-box, and the computation of the explanation is only based on its input and output [8, 11]; *intrinsic explanations* instead rely on the model itself to explain its behavior, within its algorithm [12].

Ontology Based Data Management paradigm
OBDM [13] is a paradigm based on a three level architecture, constituted by: a *source schema*, an *ontology*, and a *mapping* defining the relations between the two. This approach is a form of information integration, in which the global schema is formulated through an ontology. The ontology, however, allows us not only to reconcile different sources, but also to enrich those with symbolic knowledge from the domain of interest, which is the main reason why we resort to OBDM in this study.

Formally, an OBDM specification is a triple $J = \langle O, S, M \rangle$, where O is the ontology, S is the relational schema representing the source data, and M is a mapping from S to O. The semantics of the

OBDM system is defined by the models of J, i.e. the interpretations that satisfy both the logical theory of O and the mapping assertions in M. This implies that the semantics of the OBDM specification strongly depends on the semantic specified for M, i.e., sound, complete or exact.

3. Research Proposal

Problem Statement and Research Questions The problem at the core of our research is the lack of interpretability of many commonly used ML models. The main research question we aim to address is how to *integrate symbolic domain knowledge* with the sub-symbolic representation of data. Furthermore, we are interested in investigating the following scenarios: how to perform *semantic reasoning* tasks over sub-symbolic ML models; how symbolic reasoning can be used to investigate their behavior, check consistency, or even carry out new derived classification tasks; and lastly, how to compute *semantic explanations* to the behavior of ML models.

The Proposed Framework The key idea behind the proposed approach lies in the OBDM paradigm [13]. Therefore, the framework we aim to define is structured as a three-layer architecture: *data layer*, *ontology layer*, and *mapping*. The *data layer* and the *ontology layer* formalize, respectively, what we know on the sub-symbolic level, i.e., how is the numerical representation of the data samples defined, and what we know on the semantic level, i.e. the domain knowledge possessed by experts. However, the key innovation represented by this framework lies in the definition of a novel form of *mapping*. While mappings are well-known constructs, our approach introduces a specific form able to relate formulae expressed over the sub-symbolic, often numerical, representations of the data samples, to high level domain knowledge. This form of mapping, to the extent of our knowledge, represents a completely novel approach, which we argue is a promising technique for many of the questions presented before. In the following we provide a brief description of the framework architecture.

Data Layer
The data layer defines a formal representation of the data samples that are subject to a classification task. In ML models, data samples are usually represented as vector of *features*, also referred to as *attributes*. In our framework each *attribute* is formalized as a function, that associates to each data sample the corresponding value for that *attribute*. This formalization is crucial, as we are interested in defining a formal language to express formulae over the data samples, and therefore we need a formal representation for them. Specifically, we aim to study formulae that allow classical mathematical operators as $+, \cdot$, and comparison symbols as $<, \leq, >, \geq, =$. It can be easily verified that these are some of the operators that appear in the decision functions of many classifier models. The intuition behind this choice is that one could use this language to formalize part of, or all, the behavior of classifiers. An example of formula we aim to express with this language is $\varphi(x) = ((a_1(x)+a_2(x)) \leq 3) \vee (a_3(x) = 1)$, where a_1, a_2, a_3 are functions representing attributes, and x is a data sample for the classification task.

Ontology Layer
The role of the ontology is to formalize the knowledge possessed by experts regarding the domain of interest of the classifier. The ontology can provide both *intensional* and *extensional* knowledge to the domain. *Intensional* knowledge is expressed through logical assertions, usually referred to as the Tbox, and provide rules that should be satisfied by the data. These rules are formulated in terms of predicate symbols, defining classes and relations between individuals of the knowledge base. This approach closely mirrors human reasoning, thus facilitating understanding. Moreover, the ontology can also provide *extensional* knowledge, i.e. data populating the knowledge base, usually known in the literature as the Abox. *Extensional* knowledge can be particularly useful to inject in the domain knowledge also information about known instances, possessed by domain experts.

Mapping
A *mapping assertion* in our framework is defined as a tuple of formulae, where the first one is a formula expressed using the data language described before, and the second one is a first-order logic formula

over the ontology alphabet. For the semantics definition we envision that it might be valuable to investigate both the *sound* and the *exact* approaches. A very simple example of mapping assertion could be $\langle\{a_1(x) + a_2(x)) \leq 3\}, \{Student(x)\}\rangle$, where the first element is a formula over the attributes a_1, a_2 of a data sample x, while the second one is the concept of student in an ontology. Under the *sound* semantics this would imply that all the data samples satisfying the formula over the attributes are students. The *exact* semantics would also imply that all the students must satisfy the left formula.

Query Answering
The first task to address is query answering over the defined framework. To this end, we call *ontological specification* a tuple $J = \langle M, O\rangle$ where M is the mapping and O the ontology. Query answering in this context is tightly bounded to the concept of *certain answers*, and therefore to the definitions of *models* of J, M and O. Given a query over the ontology O, its *certain answers* with respect to the mapping M are all those data samples that are answers to that query in all the models of J. Intuitively, all those samples that satisfy both the logical theory of O and the assertions in M.

Applications This paragraphs presents a more detailed description of the tasks we aim to address within the presented framework. In particular, we focus on the two research questions that are the focus of this paper: how to compute *semantic explanations*, and how to perform and exploit *semantic reasoning* over ML models.

Formulating Explanations
As mentioned in the previous sections, an application of this framework we are interested in studying is the problem of computing explanations to the decisions of classifiers. The intuition behind the explanation definition is that it should be a formula, expressed over the ontology alphabet, such that the following properties hold: the classified data sample i is in the certain answers of that formula; and all the certain answers of that formula are assigned the same class of i by the classifier. This definition follows from the idea that an explanation should be able to represent the sample to explain, and no sample that belongs to a different class. This approach to the explanation problem falls under the category of *local*, *abductive*, *post-hoc* explanations. However, studying different definitions of the explanation is a problem we plan to address.

Regarding this specific explanation definition, we already presented some results in [14]. This paper provides a first formal definition of the framework, and studies the problem of *verifying* and *computing* an explanation for the output decision of a classifier. Being an early work, the framework is studied under several limitation assumptions. Specifically, the mapping is studied under the assumption that the left formula is quantifier-free, and uses only finite attributes, i.e., attribute functions with a finite codomain. Also, the formulae on the right-side of the mapping assertions allow only for formulae of the form $B(x)$, $\exists y.R(x, y)$, and $\exists y.R(y, x)$. And finally, the Tbox is formulated in *DL-Lite*$_R$ [15], and the Abox is empty. Moreover, we limited the study to explanations in the form of conjunctive queries. Under these assumptions, we provided algorithms for both the *verification* and *computation* problems, and studied the computational complexity of the verification, which was proved to be coNP-complete.

A first task on which we are currently working is the extension of the defined framework, with the intent of accommodating more powerful languages, and thus manage more complex problems. In particular, we are working on languages that allow the presence of constants in the ontology, as *DL-Lite*$_A$ [16]. We argue that this is a fundamental requirement to be able to express meaningful domain knowledge. It is easy to verify that this extension requires the reformulation of the entire framework in terms of many-sorted first-order logics, which should distinguish between constant values and objects to classify.

Performing Reasoning
The second task we aim to address is performing reasoning using symbolic knowledge across one or multiple ML models. A fundamental observation for the study of this tasks is that, for many popular ML models, the learned decision function can be expressed using the data language previously described. For example, Perceptrons, Support Vector Machines, Linear Regressions, all learn a decision function

in the form of a linear combination $w_1x_1 + ... + w_nx_n \geq 0$, where $w_1, ..., w_n$ are the weights, usually rational numbers, and $x_1, ..., x_n$ are the values of the data sample features. It is immediate to check that this can be expressed using the framework data language as $w_1a_1(x) + ... + w_na_n(x) \geq 0$ where $a_1, ..., a_n$ are the attribute functions that associate to the sample x its feature values $x_1, ..., x_n$.

An immediate consequence of this observation is that it could be interesting to define mapping assertions where the left formula is the decision function of a ML model, while the right-side one is the class learned by that model. For example, one might have on the left-side the decision function of a classifier for students, and on the right-side the concept $Student(x)$. These particular mapping assertions pave the way for several applications of this framework. Assume a context in which each mapping assertion represents a classifier, with its sub-symbolic decision function and symbolic ontology class. A first application of query answering over this mapping could be gaining insights on the behavior of the models, to reveal inconsistencies or undesired behaviors. Also, it could be used to synthesize new classifiers. This could be obtained by formulating queries defining a particular class, not learned by the classifiers present in the mapping. Intuitively, rewriting this query would lead to the sub-symbolic formula representing all the data samples in that class, hence a synthesized classifier for it. Lastly, another interesting application is ensemble learning. Query answering involves rewriting the query with respect to all the mapping assertions, which means that every classifier is taken into consideration in order to answer the query. This can be seen as a form of ensemble learning, where all the models participate to the classification task. Furthermore, we argue that it might be valuable to study this framework with probabilistic ontology languages, which could integrate the concept of confidence of a model, in the form of probability, or likelihood, that an individual is assigned a certain class.

4. Conclusions

The work presented in this paper introduces an innovative approach within the Neuro-symbolic AI and Explainable AI fields of research, leveraging ontologies to enhance the interpretability and transparency of ML models. Exploiting the OBDM paradigm, the presented framework allows to reconcile the sub-symbolic representation of data used by ML models, with the high-level symbolic representation of the knowledge possessed by expert users. First, we provided a description of the framework we intend to achieve, and then we presented two application scenarios for it: providing *semantic explanations*, and performing *reasoning* tasks. For the first one, we provided a summarization of the results already obtained, along with notes on its limitations and the intended future extensions. For the latter instead we described several scenarios in which classical reasoning tasks, as query answering, could be valuable in solving new interesting challenges.

In conclusion, we argue that the presented work can provide a valuable contribute to the research areas of Explainable AI and Neuro-symbolic AI, with the potential of providing novel techniques to enhance interpretability, transparency and trustworthiness of ML models.

Acknowledgments

The work presented in this paper has been supported by PNRR MUR project PE0000013-FAIR.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] M. Garnelo, M. Shanahan, Reconciling deep learning with symbolic artificial intelligence: representing objects and relations, Current Opinion in Behavioral Sciences 29 (2019) 17–23. URL:

https://www.sciencedirect.com/science/article/pii/S2352154618301943. doi:https://doi.org/10.1016/j.cobeha.2018.12.010, artificial Intelligence.

[2] H. Kautz, The Third AI Summer: AAAI Robert S. Engelmore Memorial Lecture, AI Magazine 43 (2022) 105–125. URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/19122. doi:10.1002/aaai.12036.

[3] S. Verma, J. Rubin, Fairness definitions explained, in: Proceedings of the International Workshop on Software Fairness, FairWare '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 1–7. URL: https://doi.org/10.1145/3194770.3194776. doi:10.1145/3194770.3194776.

[4] F. Zuiderveen Borgesius, Discrimination, artificial intelligence, and algorithmic decision-making, 2018. URL: https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73.

[5] European Parliament, Council of the European Union, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), 2024. URL: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32024R1689.

[6] 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. URL: http://dx.doi.org/10.1016/j.inffus.2024.102301. doi:10.1016/j.inffus.2024.102301.

[7] 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). URL: https://doi.org/10.1145/3236009. doi:10.1145/3236009.

[8] M. Cooper, L. Amgoud, Abductive Explanations of Classifiers Under Constraints: Complexity and Properties, 2023. doi:10.3233/FAIA230305.

[9] A. Darwiche, A. Hirth, On the reasons behind decisions, 2020. arXiv:2002.09284.

[10] N. Gorji, S. Rubin, Sufficient reasons for classifier decisions in the presence of domain constraints, Proceedings of the AAAI Conference on Artificial Intelligence 36 (2022) 5660–5667. URL: https://ojs.aaai.org/index.php/AAAI/article/view/20507. doi:10.1609/aaai.v36i5.20507.

[11] M. C. Cooper, J. Marques-Silva, Tractability of explaining classifier decisions, Artificial Intelligence 316 (2023) 103841. URL: https://www.sciencedirect.com/science/article/pii/S0004370222001813. doi:https://doi.org/10.1016/j.artint.2022.103841.

[12] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215.

[13] M. Lenzerini, Ontology-based data management, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM '11, Association for Computing Machinery, New York, NY, USA, 2011, p. 5–6. URL: https://doi.org/10.1145/2063576.2063582. doi:10.1145/2063576.2063582.

[14] L. Papi, G. Cima, M. Console, M. Lenzerini, Semantic explanations of classifiers through the ontology-based data management paradigm (extended abstract), in: L. G. 0001, J. C. Jung, A. Ozaki (Eds.), Proceedings of the 37th International Workshop on Description Logics (DL 2024), Bergen, Norway, June 18-21, 2024, volume 3739 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3739/abstract-22.pdf.

[15] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, R. Rosati, Tractable reasoning and efficient query answering in description logics: The *DL-Lite* family 39 (2007) 385–429.

[16] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, R. Rosati, Linking data to ontologies, in: S. Spaccapietra (Ed.), Journal on Data Semantics X, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 133–173.