Formal Reasoning with Learned Predicates (Extended Abstract)

Gianluca Cima[1], Marco Console[1] and Laura Papi[1]

[1]Sapienza University of Rome

Abstract

Machine Learning (ML) models are regularly employed in diverse application domains to support decision-making processes. In this context, the knowledge is induced by the ML models. Combining this learned knowledge with intensional knowledge specified by logic-based formalisms would be highly beneficial to these processes. In this paper we present our recent work on this subject, describing a framework that combines these two forms of knowledge and enables formal reasoning. The framework is based on the novel notion of *Hybrid Knowledge Base (HKB)*, which consists of an ontology and a set of ML binary classifiers. The ontology defines the intensional knowledge over the domain, while the ML classifiers implicitly define the extensional knowledge. Specifically, a HKB combines these two components by associating each predicate of the ontology with a classifier, which virtually populates the predicate with positively classified instances. A further contribution of our work is the study of the query answering task. In particular, we provide its computational complexity analysis for ontologies specified in (the Description Logic counterpart of) RDFS, and binary Multi-Layer Perceptrons.

Keywords

Neuro-symbolic AI, Ontology-Driven Knowledge Bases

1. Introduction

Recent advancements of Machine Learning (ML) techniques led to a widespread use of ML models. The employment of such models nowadays ranges across a variety of contexts, such as healthcare, law, finance [2, 3]. However, ML models still exhibit significant limitations that hinder their adoption in mission-critical settings [4]. One specific limiting factor lies in the interaction mechanisms available for these models. Specifically, in the case of ML classifiers, interaction is typically restricted to providing an input instance and receiving the corresponding output classification label. In this sense, the interaction with the model is *local*. However, in order to compute the classification, the model has learned a *global* knowledge over the domain. We argue that enabling access to this global acquired knowledge could be highly beneficial, although doing so is generally difficult. This paper summarizes our recent work on this subject, which introduces a novel formal framework to access and interrogate this knowledge [1]. Furthermore, this framework also allows combining this *sub-symbolic* knowledge learned by the classifiers with a *symbolic* logic-based specification of the domain of interest. Under this perspective, this approach falls into the field of Neuro-symbolic AI. The integration of such two forms of knowledge is a problem that can be studied under the lens of Ontology-Driven Knowledge Bases. The core principle of this paradigm is to represent knowledge via two complementary components: an *intensional part* and an *extensional part*. The intensional part, specified through an *ontology*, provides an intensional representation of the domain of interest. The extensional part refers to the actual data, specifying which objects (resp. pairs of objects) are instances of concepts (resp. roles). This extensional knowledge can take different forms depending on the use case [5, 6, 7]. Our novel framework introduces the notion of *Hybrid Knowledge Base (HKB)*, which models the domain through an ontology, and defines the extensional component of the knowledge base through ML classifiers. Intuitively, ML classifiers determine how the ontology concepts and roles are populated by assigning instances to predicates based on their classification labels.

The original paper has been accepted and presented at AAAI26 [1].

DL 2026: 39th International Workshop on Description Logics, July 17–19, 2026, Lisbon, Portugal

cima@diag.uniroma1.it (G. Cima); console@diag.uniroma1.it (M. Console); papi@diag.uniroma1.it (L. Papi)

0000-0003-1783-5605 (G. Cima); 0009-0004-5526-019X (M. Console); 0009-0003-2281-9500 (L. Papi)

Example 1. *Consider a medical scenario exploiting classifiers to determine whether a patient is diabetic (classifier κ_D), male (κ_M), or pregnant (κ_P). Also, assume the ontology $\mathcal{O}$ models the domain through a vocabulary containing the atomic concepts D, M, and P for diabetics, males, and pregnant patients, respectively. The background knowledge of $\mathcal{O}$ sanctions that pregnant patients cannot be male ($P \sqsubseteq \neg M$). The HKB for this example is composed by $\mathcal{O}$ and the set of classifiers described above, where κ_C is assigned to the atomic concept C (for $C \in \{D, M, P\}$).*

It is immediate to notice that the classifications provided by the classifiers can contradict some of the ontology axioms, thus leading to inconsistencies. For instance, consider the axiom $\mathsf{P} \sqsubseteq \neg\mathsf{M}$ in Example 1, and let there be a data sample simultaneously classified positively by κ_P and κ_M. Clearly, this gives rise to an inconsistency. To manage these cases we provide a model-theoretic semantics for HKBs, which specifies how to handle inconsistencies and enables formal reasoning over the combined knowledge of a HKB. The main contribution of our work focuses on the study of two reasoning tasks and the analysis of their computational complexity: *non-trivial consistency checking* and *query answering*. We study these problems in meaningful settings and consider two different semantics for query answering.

2. Framework

At the basis of this framework is the novel definition of HKB, which combines the induced knowledge of a set of binary classifiers with the intensional knowledge of the domain, modeled in terms of an ontology. As a preliminary step, it is necessary to formalize the notion of ontology and classifiers in this context.

Ontologies We define Σ_C, Σ_R, and Var to be the pairwise disjoint, countably infinite sets of atomic concepts, atomic roles, and variables, respectively. In this framework an ontology $\mathcal{O}$ consists of a finite set of declarations that use symbols from $\Sigma_C \cup \Sigma_R \cup \{=\}$ for predicates and symbols from Var for terms. We let $\mathsf{sig}(\mathcal{O})$, called the *signature* of $\mathcal{O}$, be the set of declared atomic concepts and atomic roles.

Classifiers Assume a countably infinite set $\mathbb{A}$ of *attribute symbols*, such that each $A \in \mathbb{A}$ is associated with an *attribute domain* D_A. Given a tuple $\mathcal{A} = (A_1, \ldots, A_n)$ of attributes from $\mathbb{A}$, a *binary classifier based on* $\mathcal{A}$ is a function $\kappa : \mathbb{F}(\mathcal{A}) \to \{0, 1\}$, where $\mathbb{F}(\mathcal{A}) = D_{A_1} \times \ldots \times D_{A_n}$ is the feature space of $\mathcal{A}$. Furthermore, we define a *binary classifier over pairs based on* $\mathcal{A}$ as a function $\lambda : \mathbb{F}(\mathcal{A}) \times \mathbb{F}(\mathcal{A}) \to \{0, 1\}$. The output of these functions is interpreted as a positive (resp. negative) label if the value is 1 (resp. 0).

With these notions in place, we can formalize the definition of a Hybrid Knowledge Base (HKB).

Definition 1. *A HKB $\mathcal{K}$ is a pair $\mathcal{K} = (\mathcal{O}, \Psi)$, where $\mathcal{O}$ is an ontology and Ψ is a set of classifiers based on a tuple $\mathcal{A}$ of attributes from $\mathbb{A}$ (written $\mathsf{sig}(\Psi) = \mathcal{A}$). More precisely, Ψ contains: a binary classifier for each atomic concept in $\mathsf{sig}(\mathcal{O})$; a binary classifier over pairs for each atomic role in $\mathsf{sig}(\mathcal{O})$.*

As already noted, under this definition, inconsistencies may arise when the extensional knowledge defined by the classifiers contradicts some of the ontology axioms. In order to handle these inconsistencies and enable reasoning tasks over HKBs, we turn to define their semantics, which we formalize in terms of first-order interpretations.

Definition 2. *Given a HKB $\mathcal{K} = (\mathcal{O}, \Psi)$ with $\mathsf{sig}(\Psi) = \mathcal{A}$, an* interpretation for $\mathcal{K}$ *is a pair $\mathfrak{I} = (\mathcal{I}, f)$, where: $\mathcal{I} = (\Delta^\mathcal{I}, \cdot^\mathcal{I})$ is an interpretation for $\mathcal{O}$; $f : \mathbb{F}(\mathcal{A}) \to \mathcal{P}(\Delta^\mathcal{I})$ associates to each element $\bar{a} \in \mathbb{F}(\mathcal{A})$ a (possibly empty) set $f(\bar{a})$ of objects from $\Delta^\mathcal{I}$; $f(\bar{a}) \cap f(\bar{b}) = \emptyset$ holds for every pair $(\bar{a}, \bar{b}) \in \mathbb{F}(\mathcal{A}) \times \mathbb{F}(\mathcal{A})$ such that $\bar{a} \neq \bar{b}$.*

Note that under this definition f acts as a bridge from the raw input vectors of $\mathbb{F}(\mathcal{A})$ (*sub-symbolic representations*), to the objects of the interpretation domain (*symbolic representations*). Specifically, each element $(a_1, \ldots, a_n) \in \mathbb{F}(\mathcal{A})$ represents a set of real world entities that share those same attribute values. When an interpretation associates an element $\bar{a}$ of the feature space with an empty set, it models the

fact that $\bar{a}$ represents an *infeasible* combination of values, that does not correspond to any real world entity under that interpretation. For such elements, we say they are *discarded by the interpretation*, and we write $\mathsf{disc}(\mathfrak{I}) = \{\bar{a} \in \mathbb{F}(\mathcal{A}) \mid f(\bar{a}) = \emptyset\}$. We also say that $\mathfrak{I}$ is *trivial* if $\mathsf{disc}(\mathfrak{I}) = \mathbb{F}(\mathcal{A})$.

After defining the interpretations of a HKB, we are now interested in the notion of models for HKBs.

Definition 3. *Let $\mathcal{K} = (\mathcal{O}, \Psi)$ be a HKB with $\mathsf{sig}(\Psi) = \mathcal{A}$, and let $\mathfrak{I} = (\mathcal{I}, f)$ be an interpretation for $\mathcal{K}$. We say that $\mathfrak{I}$* is a model of *$\mathcal{K}$ if $\mathcal{I} \models \mathcal{O}$ and the next conditions hold: (i) for each atomic concept $C \in \mathsf{sig}(\mathcal{O})$: $C^{\mathcal{I}} = \{o \mid \exists \bar{a}.\, \kappa_C(\bar{a}) = 1$ and $o \in f(\bar{a})\}$; (ii) for each atomic role $R \in \mathsf{sig}(\mathcal{O})$: $R^{\mathcal{I}} = \{(o, o') \mid \exists \bar{a}, \bar{b}.\, \lambda_R(\bar{a}, \bar{b}) = 1$ and $o \in f(\bar{a})$ and $o' \in f(\bar{b})\}$. When both conditions (i) and (ii) hold, we write $\mathfrak{I} \models \Psi$.*

Intuitively, a model of a HKB $\mathcal{K} = (\mathcal{O}, \Psi)$ is an interpretation that satisfies all the axioms in $\mathcal{O}$ and faithfully follows the classifiers in Ψ. The classifiers in Ψ are therefore treated as *exact mappings*, which fully determine the extensions of the ontology predicates. Under this setting, the ontology axioms act as constraints, leading to solve the inconsistencies by discarding elements of the feature space. However, it is immediate to notice that discarding elements of the feature space leads to the loss of part of the information induced by the classifiers. Therefore, a natural consequence is to restrict the focus to models that minimize the discarded information, which motivates the following definition.

Definition 4. *Given a HKB $\mathcal{K} = (\mathcal{O}, \Psi)$ and an interpretation $\mathfrak{I} = (\mathcal{I}, f)$ for $\mathcal{K}$, we say that $\mathfrak{I}$ is a* minimally-discarding model of *$\mathcal{K}$ if (i) $\mathfrak{I}$ is a model of $\mathcal{K}$ and (ii) there is no model $\mathfrak{I}' = (\mathcal{I}', f')$ of $\mathcal{K}$ such that $\mathsf{disc}(\mathfrak{I}') \subsetneq \mathsf{disc}(\mathfrak{I})$.*

Interestingly, a HKB may admit minimally-discarding models that discard different subsets of the feature space. Formally, given a HKB $\mathcal{K}$, we denote by $\mathsf{MinDisc}(\mathcal{K})$ the set of all the minimally-discarding models of $\mathcal{K}$.

2.1. Reasoning Tasks over HKBs

We investigate the problem of answering queries over HKBs, formulated using the vocabulary of the ontology. To this end, we provide the following notion of answers to queries over HKBs.

Definition 5. *Let $\mathcal{K} = (\mathcal{O}, \Psi)$ be a HKB, let $q = \{\bar{x} \mid \varphi(\bar{x})\}$ be a FOL query over $\mathcal{O}$ of arity m, and let $\bar{t} = (\bar{a}_1, \ldots, \bar{a}_m)$ be an m-tuple of elements from $\mathbb{F}(\mathcal{A})$, i.e. $\bar{t} \in \mathbb{F}(\mathcal{A})^m$. We say that $\bar{t}$ is a* skeptical-answer to q over $\mathcal{K}$ *if, for every $\mathfrak{I} = (\mathcal{I}, f) \in \mathsf{MinDisc}(\mathcal{K})$, the following condition holds: there exists an m-tuple $\bar{o} = (o_1, \ldots, o_m)$ of objects from $\Delta^{\mathcal{I}}$ such that $\mathcal{I} \models \varphi(\bar{x}/\bar{o})$ and $o_i \in f(\bar{a}_i)$, for each $i \in [m]$. For a HKB $\mathcal{K} = (\mathcal{O}, \Psi)$ and a FOL query q over $\mathcal{O}$, we let $\mathsf{Sans}(q, \mathcal{K})$ be the set of skeptical-answers to q over $\mathcal{K}$.*

In other words, in the absence of a principled criterion to single out a unique "correct" minimally-discarding model, we adopt a cautious approach by considering only those answers obtainable by all minimally-discarding models. This form of reasoning is reminiscent of classical skeptical reasoning over all repairs of a KB, a common strategy for handling query answering with respect to inconsistent KBs [8, 9].

Given the general framework presented so far, we turn our focus to the study of two fundamental reasoning tasks: *non-trivial consistency* and *skeptical entailment*. The first is the problem of checking whether a given HKB $\mathcal{K}$ admits a non-trivial model, i.e., a model $\mathfrak{I}$ such that $\mathsf{disc}(\mathfrak{I}) \subsetneq \mathbb{F}(\mathcal{A})$. The latter is the problem of, given a HKB $\mathcal{K}$, a query q , and a tuple $\bar{t}$, checking whether $\bar{t} \in \mathsf{Sans}(q, \mathcal{K})$.
For the computational complexity analysis, we assume the following scenario: the ontology $\mathcal{O}$ is specified in *DL-Lite*$^{\neg}_{\text{RDFS}}$ [10]; the classifiers in Ψ are in the class of `Multi-Layer Perceptron` (`MLP`), as defined in [11]; the query q is either in CQ or in UCQ$^{\neq}$. Furthermore, we study the problem in *classifier complexity*, i.e., with only the set Ψ of classifiers regarded as the input, while the ontology and the query are assumed to be fixed. The two following theorems present the complexity results for the two mentioned reasoning tasks.

Theorem 1. *For DL-Lite$^{\neg}_{RDFS}$ ontologies and MLP classifiers, the following holds:*

1. *Non-trivial consistency is NP-complete in classifiers complexity;*
2. *Skeptical entailment is coNExpTime-complete in classifiers complexity for queries in UCQ$^{\neq}$.*

We further observe that the hardness stated in Item 2 above holds already for queries in CQ.

In light of such a negative result, we further investigate query answering over HKBs by considering settings that are more amenable to practical solutions. To this end, we observe that for Item 2 of Theorem 1 the main source of complexity comes from the presence of axioms involving atomic roles. This naturally leads us to consider the fragment $\mathcal{H}^{\neg}$ of *DL-Lite*$^{\neg}_{\text{RDFS}}$ forbidding atomic roles in axioms.

Theorem 2. *For $\mathcal{H}^{\neg}$ ontologies, MLP classifiers, and queries in UCQ$^{\neq}$, the skeptical entailment problem becomes NP-complete in classifiers complexity. The hardness holds already for queries in CQ.*

Furthermore, we also investigate an alternative semantics for query answering, that avoids skeptically reasoning over all the minimally-discarding models. The proposed semantics is a sound approximation of the skeptical-answers semantics, and is inspired by the well-behaved *IAR semantics* adopted for query answering over inconsistent KBs [12], hence following the *WIDTIO* (*When In Doubt Throw It Out*) approach of the belief revision and update research area.

Definition 6. *We say that a model $\mathfrak{I}$ for $\mathcal{K}$ is a* minimally-discarding WIDTIO-model of $\mathcal{K}$ *if, for every $\bar{a} \in \mathbb{F}(\mathcal{A})$, the following holds: $\bar{a} \in \mathsf{disc}(\mathfrak{I})$ if and only if there exists a $\mathfrak{I}' \in \mathsf{MinDisc}(\mathcal{K})$ such that $\bar{a} \in \mathsf{disc}(\mathfrak{I}')$.*

Let now $q = \{\bar{x} \mid \varphi(\bar{x})\}$ be a query over $\mathcal{O}$ of arity m and $\bar{t} = (\bar{a}_1, \ldots, \bar{a}_m)$ be an m-tuple of elements from $\mathbb{F}(\mathcal{A})$. We say that $\bar{t}$ is a WIDTIO-answer to q over $\mathcal{K}$ *if, for every minimally-discarding WIDTIO-model $\mathfrak{I}$ of $\mathcal{K}$: there exists an m-tuple $\bar{o} = (o_1, \ldots, o_m)$ of objects from $\Delta^{\mathcal{I}}$ such that $\mathcal{I} \models \varphi(\bar{x}/\bar{o})$ and $o_i \in f(\bar{a}_i)$, for each $i \in [m]$. We denote by $\mathsf{Wans}(q, \mathcal{K})$ the set of WIDTIO-answers to q over $\mathcal{K}$.*

In other words, a WIDTIO-model implements the WIDTIO approach by discarding all and only the feature space elements discarded by at least one minimally-discarding model. Thus, the evaluation of a query under the WIDTIO semantics reasons only on those feature space elements that belong to every possible minimally-discarding model.

Similarly to the skeptical-answers semantics, we study the *WIDTIO-entailment* problem, defined as the problem of deciding, given a HKB $\mathcal{K}$, a query q, and a tuple $\bar{t}$, whether $\bar{t} \in \mathsf{Wans}(q, \mathcal{K})$. We study both the settings of *DL-Lite*$^{\neg}_{\text{RDFS}}$ and $\mathcal{H}^{\neg}$ ontologies.

Theorem 3. *For DL-Lite$^{\neg}_{RDFS}$ ontologies, MLP classifiers, and queries in UCQ$^{\neq}$, WIDTIO-entailment is Σ^p_2-complete in classifiers complexity. Under the same setting, but restricting to $\mathcal{H}^{\neg}$ ontologies, WIDTIO-entailment is NP-complete in classifiers complexity. Both hardness results already hold for CQ queries.*

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 in order to: Grammar and spelling check. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] G. Cima, M. Console, L. Papi, Foundations of formal reasoning over knowledge bases combining symbolic and sub-symbolic knowledge, Proceedings of the AAAI Conference on Artificial Intelligence 40 (2026) 18994–19002. URL: https://ojs.aaai.org/index.php/AAAI/article/view/38971. doi:10.1609/aaai.v40i23.38971.

[2] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, Y. Wang, Artificial intelligence in healthcare: past, present and future, Stroke and Vascular Neurology 2 (2017).
[3] S. Bahoo, M. Cucculelli, X. Goga, J. Mondolo, Artificial intelligence in finance: a comprehensive review through bibliometric and content analysis, SN Business & Economics 4 (2023).
[4] C. Thames, Y. Sun, A survey of artificial intelligence approaches to safety and mission-critical systems, in: 2024 Integrated Communications, Navigation and Surveillance Conference (ICNS), 2024, pp. 1–12.
[5] M. Bienvenu, M. Ortiz, Ontology-mediated query answering with data-tractable description logics, 2015, pp. 218–307.
[6] G. Xiao, L. Ding, B. Cogrel, D. Calvanese, Virtual knowledge graphs: An overview of systems and use cases, Data Intelligence 1 (2019) 201–223.
[7] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, R. Rosati, Linking data to ontologies X (2008) 133–173.
[8] D. Lembo, M. Lenzerini, R. Rosati, M. Ruzzi, D. F. Savo, Inconsistency-tolerant semantics for description logics, 2010, pp. 103–117.
[9] M. Bienvenu, C. Bourgaux, Inconsistency-tolerant querying of description logic knowledge bases, in: J. Z. Pan, D. Calvanese, T. Eiter, I. Horrocks, M. Kifer, F. Lin, Y. Zhao (Eds.), Proceedings of the Twelfth International Summer School on Reasoning Web: Logical Foundation of Knowledge Graph Construction and Query Answering (RW 2016), volume 9885 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 156–202.
[10] G. Cima, M. Console, R. M. Delfino, M. Lenzerini, A. Poggi, Answering conjunctive queries with safe negation and inequalities over RDFS knowledge bases, in: Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI 25), 2025, pp. 14824–14831.
[11] P. Barceló, M. Monet, J. Pérez, B. Subercaseaux, Model interpretability through the lens of computational complexity, in: Proceedings of the Thirty-Third Annual Conference on Neural Information Processing Systems (NeurIPS 2020), 2020.
[12] D. Lembo, M. Lenzerini, R. Rosati, M. Ruzzi, D. F. Savo, Inconsistency-tolerant query answering in ontology-based data access 33 (2015) 3–29.