Semantic Explanations of Classifiers through the Ontology-Based Data Management Paradigm (Extended Abstract)

Laura Papi, Gianluca Cima, Marco Console and Maurizio Lenzerini

Department of Computer, Control and Management Engineering, 25 Via Ariosto, Rome, 00185

Abstract

One of the main challenges in modern AI systems is to explain the decisions of complex machine learning models, and recent years have seen a burgeoning of novel approaches. These approaches often rely on some structural components of the models under consideration, e.g., the set of features used for the classification task. As a result, explanations provided by these approaches are expressed in terms of the sub-symbolic information and, therefore, they are hard to interpret for users. In this paper, we argue that, in order to foster interpretability, these explanations should be expressed in terms of the knowledge that the users posses on the underlying application domain rather than on the sub-symbolic components of the model. To this end, our first contribution is the illustration of a novel formal framework for explaining the decisions of machine learning classifiers grounded on the Ontology-Based Data Management paradigm. Within this framework, explanations are defined by logical formulae using the symbols that an ontology defines and, as such, they posses a well-defined semantics. As a second contribution, we provide an algorithm that computes the best explanations that can be expressed in the class of conjunctive queries.

Keywords

Ontology-Based Data Management, Machine Learning Classifiers, Explainable AI.

1. Introduction

Classifiers form a prominent family of modern AI systems. Intuitively, a classifier is a systems used to predict whether an object belongs to a specific class given a set of its relevant attributes [1]. Due to the nature of the techniques involved, the behavior of classifiers is often regarded as opaque by end users [2] and several techniques have been proposed to elucidate it [2]. An important notion in this context is that of *local explanations*, i.e., answers for the question *why a given object is assigned to a specific class*. Concretely, these explanations usually consist of a set of properties of the given object that dictate the behavior of the classifier expressed in terms of the *raw data attributes* used to operate it [3, 4, 5, 6].

While explanations based on raw data attributes may convey some information to AI Experts, it is often hard for general users to understand their meaning. This is especially true in the typical machine learning scenario where attributes are the results of a complex process of feature selection and carry little to no meaning by themselves. The goal of our work is to define a novel framework to express explanations using *conceptual properties of the scenario of interest* that are not limited by the data attributes used by the classifier.

Our framework is based on the notion of *mappings*, well-known by the AI community and widely used in the context of Information Integration [7] and Ontology-Based Data Management [8]. These mappings define the relation between the objects of the world that are relevant for a classifier and a set of conceptual notions that are relevant for the application domain. To formalize these conceptual notions, our framework makes use of ontologies that formalize the application domain. Combining domain ontologies and mappings is a well-established approach to lift information about raw data to the

✉ laup.97@gmail.com (L. Papi); cima@diag.uniroma1.it (G. Cima); console@diag.uniroma1.it (M. Console); lenzerini@diag.uniroma1.it (M. Lenzerini)

🆔 0009-0003-2281-9500 (L. Papi); 0000-0003-1783-5605 (G. Cima); 0009-0004-5526-019X (M. Console); 0000-0003-2875-6187 (M. Lenzerini)

conceptual level [9, 10, 11, 12]. In our framework, these combinations, called *ontological specifications*, are used to formalize the relation between the classifier whose behavior we want to explain and the notions that users understand. We then use ontological specifications to provide a *local explanation of a classifier expressed at the conceptual level of their ontologies via their mappings*. In this way, we obtain explanations expressed as logical formulae over the symbols of the ontology and grounded on a formal semantics.

In this context, the contribution of this paper is the following. Firstly, we present the framework of ontological specifications together with a suitable notion of explanation (Section 2). Secondly, when ontologies and mappings are expressed in reasonably expressive languages, we study the computational complexity of verifying whether a given formula is an explanation. Finally, we present a general algorithm for the computation of best explanations (Section 3).

2. Formal Framework

We proceed to present our framework for semantic explanations of ML models. Assume a possibly infinite set Δ of elements that we call *instances*. Intuitively, Δ is the set of all possible elements that the instance space of an ML model in our framework may possibly contain. Observe that instances are not yet characterized by their attributes as it is customary in learning algorithms. To bridge this gap, we further assume a countably infinite set $\mathbb{A}$ of unary function symbols that we call the set of *attribute symbols*. To each $a_i \in \mathbb{A}$, we associate a *surjective function* $a_i^{\mathsf{sem}} : \Delta \to \mathcal{D}_i$ that we call the *semantics of a_i*. Whenever the co-domain of a_i^{sem} is finite, we say that a_i is a *finite attribute*. Intuitively, a pair $\mathcal{K} = \langle \Delta, \mathbb{A} \rangle$ provides a formal background to instance space elements and, for this reason, we refer to it as a *data layer*.

A *classifier* for Δ is a function γ from Δ to $\{0, 1\}$. Usually, classifiers operate on a restricted set of attributes of the input instances. To capture this property, we say that a classifier γ *operates over a set of attributes* $A \subseteq \mathbb{A}$ if, for every pair $d, d' \in \Delta$, the fact that $a_i(d) = a_i(d')$, for each $a_i \in A$, implies $\gamma(d) = \gamma(d')$. We will call *A relevant attributes* for γ. A classifier γ for Δ is a *$\mathcal{K}$-classifier* if there exists a unique and finite set of relevant attributes $A \subseteq \mathbb{A}$ for γ.

Let $\mathcal{D}$ be the set of all possible values that an attribute in $\mathbb{A}$ may take, i.e., $\mathcal{D} = \bigcup_i \mathcal{D}_i$. We assume two countably infinite sets $\mathbb{F}$ and $\mathbb{C}$ of *function symbols* and *relation symbols*, respectively. For each $f_i \in \mathbb{F}$ with arity n, the *semantics of f_i* is a function $f_i^{\mathsf{sem}} : \mathcal{D}^n \to \mathcal{D}$. Similarly, for each $R_i \in \mathbb{C}$ with arity n, the *semantics of R_i* is a relation $R_i^{\mathsf{sem}} \subseteq \mathcal{D}^n$. Intuitively, $\mathbb{A}$, $\mathbb{F}$, and $\mathbb{C}$ will form the terms of our declarative language. Assume a countably infinite set of variables $\mathcal{V}$, the set $Terms_{\mathcal{K}}(\mathbb{F}, \mathbb{C})$ is the set of all the expressions of the following forms: d, with $d \in \Delta$, $a(x)$, with $a \in \mathbb{A}$ and $x \in \mathcal{V}$, or $f(t_1, \ldots, t_n)$, with f a function symbol of arity n in $\mathbb{F}$ and $t_1, \ldots, t_n \in Terms_{\mathcal{K}}(\mathbb{F}, \mathbb{C})$. The language $\mathcal{L}_{\mathcal{K}}(\mathbb{F}, \mathbb{C})$ is defined as the set of all first-order formulae that can be expressed using terms in $Terms_{\mathcal{K}}(\mathbb{F}, \mathbb{C})$. The semantics of $\mathcal{L}_{\mathcal{K}}(\mathbb{F}, \mathbb{C})$ is defined as customary using d, a^{sem}, and f^{sem} to interpret $d \in \Delta$, $a \in \mathbb{A}$ and $f \in \mathbb{F}$, respectively. Given $\varphi \in \mathcal{L}_{\mathcal{K}}(\mathbb{F}, \mathbb{C})$ with free variables $\bar{x}$ and a function $v : \mathcal{V} \to \Delta$, we write $v \models \varphi$ to say that the formula obtained from φ by replacing each $x \in \bar{x}$ with $v(x)$ is true.

Assume a countably infinite set of predicate symbols $\mathbb{P}$ disjoint from $\mathbb{C}$. A *mapping assertion* from $\mathcal{L}_{\mathcal{K}}(\mathbb{F}, \mathbb{C})$ to $\mathbb{P}$ is an expression of the form $\langle \varphi(x), \psi(x) \rangle$ where $\varphi(x)$ is a formula in $\mathcal{L}_{\mathcal{K}}(\mathbb{F}, \mathbb{C})$ with one free variable x and $\psi(x)$ is a first-order formula over $\mathbb{P}$ with the single free variable x. A *mapping* from $\mathcal{L}_{\mathcal{K}}(\mathbb{F}, \mathbb{C})$ to $\mathbb{P}$ is a finite set of mapping assertions from $\mathcal{L}_{\mathcal{K}}(\mathbb{F}, \mathbb{C})$ to $\mathbb{P}$. Intuitively, mappings define the connection between the instances in the data layer and the predicates in $\mathbb{P}$. To express such connection, we use *ontological specifications*.

Formally, an *ontological specification* for $\mathcal{L}_{\mathcal{K}}(\mathbb{F}, \mathbb{C})$ to $\mathbb{P}$ (simply, ontological specification) is a pair $\mathcal{O} = \langle M, T \rangle$ where T is a first-order theory over $\mathbb{P}$ and M is a mapping from $\mathcal{L}_{\mathcal{K}}(\mathbb{F}, \mathbb{C})$ to $\mathbb{P}$. The semantics of an ontological specification is defined in terms of its models. An *interpretation for $\mathbb{P}$* (simply, interpretation) is a first-order logic interpretation $\mathcal{I}$ for the symbols of $\mathbb{P}$ whose domain is Δ. Given a mapping assertion $m = \langle \varphi, \psi \rangle$, we say that $\mathcal{I}$ *satisfies m*, if, for every function $v : \mathcal{V} \to \Delta$, $v \models \varphi$ implies $v, \mathcal{I} \models \psi$. A *model for $\mathcal{O}$* is an interpretation $\mathcal{I}$ such that $\mathcal{I}$ satisfies T and $\mathcal{I}$ satisfies m, for

each $m \in M$. We use $mod(\mathcal{O})$ for the set of all models of $\mathcal{I}$.

With ontological specifications in place, we are now ready to formalize our notion of explanation. Let $\mathcal{O}$ be an ontological specification as above and φ a first-order formula over $\mathbb{P}$. We use $cert(\varphi, \mathcal{O})$ for the set $\{j \in \Delta \mid j \in \varphi^{\mathcal{I}}, \text{ for each } \mathcal{I} \in mod(\mathcal{O})\}$. Assume now a classifier γ for the data layer $\mathcal{K}$ and an instance $i \in \Delta$. A *Weak Ontology-Based eXplanations (w-OBX) for the decision of γ over i based on $\mathcal{O}$* is a first-order formula $\eta(x)$ over the alphabet $\mathbb{P}$ and one free variable x with the following properties: $i \in cert(\eta, \mathcal{O})$, and, $\gamma(i) = \gamma(j)$, for each $j \in cert(\eta, \mathcal{O})$. The next definition formalizes the notion of explanation we are looking for.

Definition 1. *Let L be a language of first-order formulae over $\mathbb{P}$. A w-OBX η for the decision of γ over i based on $\mathcal{O}$ is the best Ontology-Based Explanation in L (L-OBX) if $\eta \in L$ and there exists no w-OBX η' for the decision of γ over i based on $\mathcal{O}$ such that $\eta' \in L$ and $cert(\eta, \mathcal{O}) \subsetneq cert(\eta', \mathcal{O})$.*

Example 1. *Consider a scenario where a classifier γ is used to provide movie recommendations. The relevant attributes for γ are cr (Critic Rating) and pr (Public Rating) with domain $[0, 10]$; and lb (Low Budget) and fc (Famous Cast) with domain $\{y, n\}$. Moreover, $\gamma(i) = 1$ if and only if i satisfies the following $\mathcal{L}_{\mathcal{K}}(\mathbb{F}, \mathbb{C})$ formula: $\left(\left(\frac{1}{2} \cdot (pr(x) + cr(x)) \right) \geq 5 \right) \wedge (fc(x) = n)$. Intuitively, γ recommends a movie if it received a good average score from critics and public and it stars non-famous actors. Suppose that we want to explain the decision $\gamma(i) = 1$ taken by γ for the movie i such that $pr(i) = 10$, $cr(i) = 10$, $lb(i) = yes$, $fc(i) = no$. For the explanation, we want to use the ontological symbols PA (Publicly Acclaimed), CA (Critically Acclaimed), BM (B-Movie), and CM (Cult Movie). Let T and M be, respectively, the TBox $\{PA \sqsubseteq CM; CA \sqsubseteq CM; \}$ and mapping $\{m_1, m_2, m_3\}$ with $m_1 = \{(pr(x) = 10), PA(x)\}$, $m_2 = \{(cr(x) = 10), CA(x); m_3 = \{((lb(x) = yes) \wedge (fc(x) = no)), BM(x)\}$. Let $\mathcal{O} = \langle M, T \rangle$. It is easy to verify that the following are all w-OBX for the decision of γ over i based on $\mathcal{O}$: $(PA(x) \wedge BM(x))$, $(CA(x) \wedge BM(x))$, and $(CM(x) \wedge BM(x))$. However, $CM(x) \wedge BM(x)$ is the only CQ-OBX for the decision of γ over i based on $\mathcal{O}$, where CQ is the language of conjunctive queries.*

3. Some Preliminary Technical Results

Let $\mathcal{L}_{\mathcal{K}}^{-}(\mathbb{F}, \mathbb{C})$ be the quantifier-free subset of $\mathcal{L}_{\mathcal{K}}(\mathbb{F}, \mathbb{C})$ that uses only finite attributes. In what follows, we assume that $i)$ classifiers and formulae $\varphi(x)$ in the left-hand side of mapping assertions are defined in $\mathcal{L}_{\mathcal{K}}^{-}(\mathbb{F}, \mathbb{C})$; $ii)$ the right-hand side of mapping assertions allows only for formulae of the form $B(x)$, $\exists y. R(x, y)$, and $\exists y. R(y, x)$; $iii)$ theories over $\mathbb{P}$ are formulated in *DL-Lite$_{\mathcal{R}}$* [13]; and $iv)$ the language for expressing explanations is the class of conjunctive queries CQ. In this scenario, we consider the following computational problems. **Verification:** given also a CQ $\eta(x)$ over the alphabet $\mathbb{P}$, check whether η is a w-OBX of the decision of γ over i based on $\mathcal{O}$; **Computation:** compute all the CQ-OBXs of the decision of γ over i based on $\mathcal{O}$.

Theorem 1. *Verification is coNP-complete.*

Next, we provide a technique to return the set of all CQ-OBXs of the decision of γ over i based on $\mathcal{O}$ (clearly, if two formulae $q(x)$ and $q'(x)$ are such that $cert(q, \mathcal{O}) = cert(q', \mathcal{O})$, then we say that they are equivalent w.r.t. $\mathcal{O}$ and treat them as the same formula).

Given an instance $i \in \Delta$ and a mapping M from $\mathcal{L}_{\mathcal{K}}(\mathbb{F}, \mathbb{C})$ to $\mathbb{P}$ in our considered scenario, we denote by $M(i)$ the set of atoms obtained by *chasing* the instance i w.r.t. M, i.e: $M(i)$ contains the atom $B(i)$ (resp. $\exists R(i), \exists R^{-}(i)$) if and only if there exists a mapping assertion of the form $\langle \varphi(x), B(x) \rangle$ (resp. $\langle \varphi(x), \exists y. R(x, y) \rangle$, $\langle \varphi(x), \exists y. R(y, x) \rangle$) in M such that $\varphi(i)$ is true. Furthermore, given a set $M(i)$ of atoms as above, we denote by $\eta_M^i(x)$ the CQ obtained by conjoining all the atoms in $M(i)$, where we select a free variable x and each atom of the form $B(i)$ is replaced with $B(x)$, and each atom of the form $\exists R(i)$ (resp. $\exists R^{-}(i)$) is replaced with $\exists y. R(x, y)$ (resp. $\exists y. R(y, x)$) in which y is always a fresh existential variable. Given an instance $i \in \Delta$ and an ontology $\mathcal{O} = \langle M, T \rangle$ in our scenario, we now prove that $\eta_M^i(x)$ is actually the *smallest (up to equivalence w.r.t. $\mathcal{O}$)* CQ such that $i \in cert(\eta_M^i, \mathcal{O})$,

in the sense that there exists no other CQ $q(x)$ for which $i \in cert(q, \mathcal{O})$ and there is an instance $j \in \Delta$ satisfying $j \in cert(\eta_M^i, \mathcal{O})$ and $j \notin cert(q, \mathcal{O})$.

Proposition 1. *Given an instance $i \in \Delta$ and an ontology $\mathcal{O} = \langle M, T \rangle$, we have that $\eta_M^i(x)$ is the smallest (up to equivalence w.r.t. $\mathcal{O}$) CQ such that $i \in cert(\eta_M^i, \mathcal{O})$.*

Given an instance $i \in \Delta$ and an ontology $\mathcal{O} = \langle M, T \rangle$ in our considered scenario, we denote by $M_{\mathcal{O}}(i)$ the set of atoms obtained from $M(i)$ by adding the atom $C(i)$ if and only if there exists an atom of the form $C'(i) \in M(i)$ and $T \models C' \sqsubseteq C$, where both C and C' can be any basic *DL-Lite$_{\mathcal{R}}$* concept, i.e. concepts of the form B, $\exists R$, and $\exists R^-$ with B and R in $\mathbb{P}$.

Theorem 2. *Let γ be a classifier, $i \in \Delta$ be an instance, $\mathcal{O} = \langle M, T \rangle$ be an ontology specification, and $\eta(x)$ be a CQ-OBX of the decision γ over i w.r.t. $\mathcal{O}$. We have that $\eta(x)$ is equivalent w.r.t. $\mathcal{O}$ to a query of the form $\eta_{M'}^i(x)$, where $M' \subseteq M_{\mathcal{O}}(i)$.*

Actually, the above results suggest a naive algorithm to compute the set of all the CQ-OBXs. Specifically, it is enough to consider all the possible $\eta_{M'}^i(x)$, where $M' \subseteq M_{\mathcal{O}}(i)$, and check that 1) $\eta_{M'}^i(x)$ is a w-OBX of the decision γ over i based on $\mathcal{O}$, and 2) there is no other $M'' \subseteq M_{\mathcal{O}}(i)$ for which $\eta_{M''}^i(x)$ is a w-OBX of the decision γ over i based on $\mathcal{O}$ and the formula PERFECTREF$(\eta_{M'}^i, \mathcal{O})$ is strictly contained in the formula PERFECTREF$(\eta_{M''}^i, \mathcal{O})$, meaning that it can be the case that $cert(\eta_{M'}^i, \mathcal{O}) \subsetneq (\eta_{M''}^i, \mathcal{O})$. Here, PERFECTREF denotes the algorithm used for rewriting CQs w.r.t. *DL-Lite$_{\mathcal{R}}$* TBoxes [13].

Acknowledgments

References

[1] S. Shalev-Shwartz, S. Ben-David, Understanding Machine Learning - From Theory to Algorithms, Cambridge University Press, 2014. URL: http://www.cambridge.org/de/academic/subjects/computer-science/pattern-recognition-and-machine-learning/understanding-machine-learning-theory-algorithms.

[2] A. B. Arrieta, N. D. Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115. URL: https://doi.org/10.1016/j.inffus.2019.12.012. doi:10.1016/J.INFFUS.2019.12.012.

[3] M. C. Cooper, J. Marques-Silva, Tractability of explaining classifier decisions, Artif. Intell. 316 (2023) 103841. URL: https://doi.org/10.1016/j.artint.2022.103841. doi:10.1016/J.ARTINT.2022.103841.

[4] A. Shih, A. Choi, A. Darwiche, A symbolic approach to explaining bayesian network classifiers, in: J. Lang (Ed.), Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, ijcai.org, 2018, pp. 5103–5111. URL: https://doi.org/10.24963/ijcai.2018/708. doi:10.24963/IJCAI.2018/708.

[5] A. Darwiche, Three modern roles for logic in AI, in: D. Suciu, Y. Tao, Z. Wei (Eds.), Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2020, Portland, OR, USA, June 14-19, 2020, ACM, 2020, pp. 229–243. URL: https://doi.org/10.1145/3375395.3389131. doi:10.1145/3375395.3389131.

[6] Y. Izza, J. Marques-Silva, On explaining random forests with SAT, in: Z. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021, ijcai.org, 2021, pp. 2584–2591. URL: https://doi.org/10.24963/ijcai.2021/356. doi:10.24963/IJCAI.2021/356.

[7] M. Lenzerini, Data integration: A theoretical perspective., in: Proceedings of the Twenty-First ACM SIGACT SIGMOD SIGART Symposium on Principles of Database Systems (PODS 2002), 2002, pp. 233–246.

[8] M. Lenzerini, Ontology-based data management, in: Proceedings of the Twentieth International Conference on Information and Knowledge Management (CIKM 2011), 2011, pp. 5–6. doi:10.1145/2063576.2063582.

[9] G. Cima, M. Console, M. Lenzerini, A. Poggi, A review of data abstraction, Frontiers Artif. Intell. 6 (2023). URL: https://doi.org/10.3389/frai.2023.1085754. doi:10.3389/FRAI.2023.1085754.

[10] F. Croce, G. Cima, M. Lenzerini, T. Catarci, Ontology-based explanation of classifiers, in: A. Poulovassilis, D. Auber, N. Bikakis, P. K. Chrysanthis, G. Papastefanatos, M. A. Sharaf, N. Pelekis, C. Renso, Y. Theodoridis, K. Zeitouni, T. Cerquitelli, S. Chiusano, G. Vargas-Solar, B. Omidvar-Tehrani, K. Morik, J. Renders, D. Firmani, L. Tanca, D. Mottin, M. Lissandrini, Y. Velegrakis (Eds.), Proceedings of the Workshops of the EDBT/ICDT 2020 Joint Conference, Copenhagen, Denmark, March 30, 2020, volume 2578 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2578/PIE3.pdf.

[11] T. Catarci, M. Scannapieco, M. Console, C. Demetrescu, My (fair) big data, in: J. Nie, Z. Obradovic, T. Suzumura, R. Ghosh, R. Nambiar, C. Wang, H. Zang, R. Baeza-Yates, X. Hu, J. Kepner, A. Cuzzocrea, J. Tang, M. Toyoda (Eds.), 2017 IEEE International Conference on Big Data (IEEE BigData 2017), Boston, MA, USA, December 11-14, 2017, IEEE Computer Society, 2017, pp. 2974–2979. URL: https://doi.org/10.1109/BigData.2017.8258267. doi:10.1109/BIGDATA.2017.8258267.

[12] G. Cima, A. Poggi, M. Lenzerini, The notion of abstraction in ontology-based data management, Artificial Intelligence 323 (2023) 103976.

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