Evaluating the Interpretability of Tooth Expressions

Extended Abstract

Guendalina Righetti[1], Daniele Porello[2] and Roberto Confalonieri[1]

[1]Free University of Bozen-Bolzano, Faculty of Computer Science, 39100, Bolzano, Italy
[2]Università di Genova, Dipartimento di Antichità, Filosofia e Storia, 16126, Genova, Italy

Abstract

In Knowledge Representation, Tooth expressions have been shown to behave like linear classification models. Thus, they can provide a powerful yet natural tool to represent local explanations of black box classifiers in the context of Explainable AI. In this extended abstract, we present the result of a user study in which we evaluated the interpretability of Tooth expressions compared to Disjunctive Normal Forms (DNF). In the user study, we asked respondents to carry out two classification tasks using concepts represented either as Tooth expressions or as different types of DNF formulas. We evaluated interpretability through accuracy, response time, confidence, and perceived understandability by human users. In line with our hypothesis, the study revealed that Tooth expressions are generally faster to use, and that they are perceived more understandable by users who are less familiar with logic.

Keywords

Tooth expressions, Explainable AI, Interpretability, User study

1. Introduction

Symbolic knowledge plays a key role for the creation of intelligible explanations. In [1], it has been shown that the integration of DL ontologies in the creation of explanations can enhance the perceived *interpretability*[1] of post-hoc explanations by human users.

Motivated by the conventional wisdom that disjunctive normal form (DNF) is considered as a benchmark in terms of both expressivity and interpretability of logic-based knowledge representations [3], we assume to have local explanations of black box models modeled as a DNF formula. An example explanation from a loan agent could be: 'I grant a loan when the subject has no children and is married or when he has high income range' (i.e., $(\neg Parent \sqcap Married) \sqcup Rich$). Prior works raised the questions of whether DNF is always the most interpretable representation, and whether alternate representation forms enable better interpretability [3, 4]. In particular, [4] evaluated several forms of DNFs in terms of their interpretability when presented to human users as logical explanations for different domains of application. In this work we aim at comparing the intepretability of DNFs and Tooth expressions.

DL 2022: 35th International Workshop on Description Logics, August 7–10, 2022, Haifa, Israel

EMAIL: Guendalina.Righetti@stud-inf.unibz.it (Guendalina Righetti); daniele.porello@unige.it (Daniele Porello); roberto.confalonieri@unibz.it (Roberto Confalonieri)

[1] Interpretability describes the possibility to comprehend a black box model and to present the underlying basis for decision-making in a way that is understandable to humans [2].

Tooth expressions have been studied in the context of Knowledge Representation and integrated within DLs in [5], by adding a novel concept constructor, the "Tooth" operator ($\mathbb{W}$). Tooth operators allow for introducing weights into standard DL languages to assess the importance of the features in the definition of the concepts. For instance, as we shall see, the concept $\mathbb{W}^1((Parent, -1), (Rich, 2), (Married, 1))$ classifies those instances for which the sum of the satisfied weighted concepts reaches the threshold 1.

In the context of XAI, Tooth expressions provide a powerful yet natural tool to represent local explanations of black box classifiers. In [6, 7] a link between Tooth-expressions and linear classifiers has been established, where it is shown that Tooth-operators behave like *perceptrons*. Furthermore, adding Tooth operators to any language including the booleans does not increase the expressivity and complexity of the language. Tooth expressions are indeed equivalent to standard DNFs[2] [7]: they are 'syntactic sugar' for languages that include the booleans. They allow, however, for crisper formulas, being thus less error-prone and, putatively, more understandable. Moreover, the representation of Tooth expressions is inspired by the design of Prototype Theory [8]. Tooth operators are thus more cognitively grounded than standard logic languages, allowing for a representation of concepts that is, arguably, more in line with the way humans think of them [9, 10].

In this paper, we present the results of a user study we conducted to measure the interpretability of Tooth expressions versus their translation into standard DNFs. In the user study, respondents were asked to carry out different classification tasks using concepts represented both as a Tooth-expressions and as DNFs. In line with previous works evaluating the interpretability of explanation formats (e.g., [11, 1, 4, 12, 13]), we used the metrics of accuracy, time of response, and confidence in the answers as a proxy for evaluating the interpretability of the two representations. We expected that Tooth expressions could be perceived more interpretable. In line with our hypothesis, our study revealed that the type of task, the background of the respondents, and the size of the DNF formula affect the interpretability of the formalism used.

2. Background

Tooth Operator - Preliminary Definitions. In this section, we delineate the formal framework necessary to introduce $\mathbb{W}$ (Tooth) expressions. Following the work done in [5], we extend standard DL languages with a class of m-ary operators denoted by the symbol $\mathbb{W}$ (spoken 'Tooth'). Each operator works as follows: (i) it takes a list of concepts, (ii) it associates a weight (i.e., a number) to each of them, and (iii) it returns a complex concept that applies to those instances that satisfy a certain combination of concepts, i.e., those instances for which, by summing up the weights of the satisfied concepts, a certain threshold is met. More precisely, we assume a vector of m weights $\vec{w} \in \mathbb{R}^m$ and a threshold value $t \in \mathbb{R}$. If $C_1, \dots, C_m$ are concepts of $\mathcal{ALC}$, then $\mathbb{W}^t_{\vec{w}}(C_1, \dots, C_m)$ is a concept of $\mathcal{ALC}_{\mathbb{W}}$. To better visualise the weights an operator associates to the concepts, we often use the notation $\mathbb{W}^t((C_1, w_1), \dots, (C_m, w_m))$ instead of $\mathbb{W}^t_{\vec{w}}(C_1, \dots, C_m)$. The semantics of $\mathcal{ALC}_{\mathbb{W}}$ just extends the usual semantics of $\mathcal{ALC}$ to account for the interpretation of the Tooth operator, as follows. Let $I = (\Delta^I, \cdot^I)$ be an

[2] More precisely, non-nested Tooth-expressions are not able to represent the XOR. Nested Tooth can however overcome this difficulty.

interpretation of $\mathcal{ALC}$. The interpretation of a $\mathbb{W}$-concept $C = \mathbb{W}^t((C_1, w_1), \ldots, (C_m, w_m))$ is: $C^I = \{d \in \Delta^I \mid v_C^I(d) \geq t\}$ where $v_C^I(d)$ is the *value* of $d \in \Delta^I$ under the concept C, i.e.: $v_C^I(d) = \sum_{i \in \{1,\ldots,m\}} \{w_i \mid d \in C_i^I\}$.

We refer the interested reader to [5, 6, 14] for a more precise account of the properties of the operator.

Disjunctive Normal Forms - Preliminary Definitions. A disjunctive normal form (DNF) is a logical formula consisting of a disjunction of one or more conjunctions, of one or more literals. In our study, we used DL symbols (⊓, ⊔) to interpret conjunctions and disjunctions.

We will follow the definitions from [3]. Accordingly, DNF is a strict subset of the Negation Normal Form language. An NNF formula is characterised as a *rooted*, *directed*, *acyclic* graph, where each leaf node is labeled with a propositional variable or its negation, and each internal node is labeled with a conjunction or a disjunction. A DNF is a *flat* NNF, i.e., an NNF whose maximum number of edges from the root to some leaf is 2. Moreover, DNFs satisfies the property of *simple conjunction*, i.e., each propositional variable occurs at most once in each conjunction.

One can consider different NNF subsets by imposing one or more of the following conditions on the formulas: (i) *Decomposability*: an NNF is decomposable (DNNF) iff for each conjunction in the NNF, the conjuncts do not share variables. Each DNF is decomposable by definition. (ii) *Determinism*: an NNF is deterministic (d-NNF) iff for each disjunction in the NNF, every two disjuncts are logically contradictory. (iii) *Smoothness*: NNFs satisfy smoothness (sd-NNF) iff for each disjunction formula, each disjunct mentions the same variables. When looking at DNF, the class of formulas satisfying determinism and smoothness is called MODS.

In what follows, we will consider three sets of DNF, obtained by adding different conditions on the formulae (and leading to formulas of different sizes): (i) **DNF1**: Simple (decomposable) DNFs ($DNF1 \subsetneq DNNF$), corresponding to the shorter formulas. The only requirement for the formulas is to satisfy the property of simple conjunction. (ii) **DNF2**: Deterministic DNFs ($DNF2 \subsetneq d - NNF$), for which each couple of disjuncts is required to be logically contradictory. (iii) **DNF3**: Deterministic, smooth DNFs ($DNF3 \subsetneq MODS$), corresponding to the longest possible DNFs. DNF3 collect all the formula models.

3. Experimental Evaluation

Method. We had 6 concept definitions of different complexities. For each concept, we constructed four formulas, one for each of the formats (DNF1, DNF2, DNF3 and Tooth expression). We thus obtained 24 distinct concept definitions. Each participant was shown twelve formulas corresponding to concept definitions, randomly shuffled.

We had two distinct online questionnaires, one for the DNFs and one for Tooth expressions. The questionnaire was run in a controlled environment (i.e., in a classroom) and contained an introductory and an experimental phase. In the introductory phase, subjects were shown a short description of either DNFs or Tooth operator, and how its semantics is determined. The experimental phase was subdivided into two tasks: classification, and inspection. In these tasks the participants were presented with six formulas corresponding to one of the two representations. In the classification task, subjects were asked to decide if a certain combination

Table 1
Mean values of correct answers, time of response, user confidence, and user understandability for formulas represented using DNFs and Tooth operator for **GroupI** and **GroupII** (standard deviations are reported in parenthesis).

Group	Measure	DNFs	Tooth
Computer Science	%Correct Responses	0.90 (0.32)	0.88 (0.31)
	Time (sec)	37.29 (55.29)	25.23 (17.96)
	Confidence	5.98 (1.29)	5.73 (1.71)
	User Understandability	6.11 (1.17)	5.61 (1.65)
Philosophy	%Correct Responses	0.86 (0.30)	0.90 (0.34)
	Time (sec)	36.39 (28.06)	24.80 (16.77)
	Confidence	5.44 (1.28)	5.88 (1.15)
	User Understandability	5.43 (1.20)	5.84 (1.10)

of literals is an instance of a given formula (e.g., *Given the formula* $C_1 := (\neg A \sqcap C) \sqcup B$*. If* i *is* $\neg A$*,* B*, and* $\neg C$*, then* i *is an instance of* C_1). In the inspection task, participants had to decide on the truth value of a particular statement, referring to if some given conditions of an instance are necessary for the instance to belong to a given class (e.g., *Given the formula* $C_1 := (\neg A \sqcap C) \sqcup B$*. Having* B *is necessary for being classified as* C_1). While the former task provides all details necessary for performing the decision, the latter only specifies whether a subset of the features influence the decision. In these two tasks, for each formula, we recorded: (i) Correctness of the response; (ii) Confidence in the response, as provided on a Likert scale from 1 to 7; (iii) Response time measured from the moment the formula was presented; (iv) Perceived formula understandability, as provided on a Likert scale from 1 to 7.

58 participants volunteered to take part in the experiment. We had two groups of students, 33 students with a background in computer science and 25 students with a background in philosophy. Each group repeated the questionnaire twice, once using DNFs and once using Tooth expressions. In the analysis, we will denote these groups as GroupI and GroupII respectively.

Results. When looking at the two groups together, we observed a significant influence ($p < .0001$) of Tooth expressions on the time of response within both tasks, showing that when using Tooth operators respondents carried out the tasks in a quicker way. This suggests that Tooth expressions are more cognitively friendly than standard DNFs.

When looking at the two groups separately (Table 1), the percentages of correct answers are slightly different when using DNFs and Tooth operators, but this difference is not statistically significant. Thus, we can conclude that the type of formula used does not have any significant effects or interactions on the accuracy of responses. Tooth operators yielded faster responses in both groups. This seems to suggest that having more compact information, like in the case of Tooth operators, could speed up the human decision-making process. Interestingly, faster decision making can yield more correct responses, but surprisingly faster decision-making is not always associated with highest perceived understandability and highest confidence. Respondents with computer science background were more confident with DNFs and perceived them as more understandable than Tooth operators. On the contrary, respondents with a background in philosophy found Tooth operators more understandable and were more confident with their answers when using Tooth operators. This behaviour can be motivated by the fact that computer scientists were introduced to logic and DNF formulas in their curricula, but not to Tooth operators. Thus, being more proficient in DNFs, they did not face the 'learning curve' in understanding a new representation formalism such Tooth operators. Respondents with a background in philosophy, on the other hand, studied neither DNFs nor Tooth operators. From this study, we can conclude that Tooth operators are better representation for users who are not familiar with logic, and with DNFs in particular.

When looking at results of different DNFs vs Tooth operator we observed that simpler DNF formats, namely DNF1 and DNF2, yielded more accurate responses. Tooth operators perform better compared to DNF3. This is expected since formulas in DNF3 format tend to be very long. DNF1 and DNF2 performed similarly in our study. This is expected, since they are quite similar in lengths. As far as time is concerned, we still observe that Tooth operators are faster than any of the DNF formats. The 'interformat' analysis seems to suggest that DNF1 and Tooth operator have quite similar understandability from the performance point of view and also from the subjective point of view. On the other hand, DFN2 and DNF3 require longer time of response and were perceived less understandable than Tooth operators.

4. Conclusion and Future Works

In this paper, we compared the intepretability of Tooth expressions and a standard logical formalism, i.e., the DNFs through a user study. In the study, we asked users to carry out two distinct tasks, namely classification and inspection (see Section 3), using Tooth expressions and DNFs. The interpretability of Tooth expressions and DNFs was measured through human-grounded metrics [2], namely accuracy in the responses, time of response, confidence in the responses, and perceived understandability.

We observed that Tooth expressions were generally faster to process, leading to a lower time of response. This was observed across all different DNFs formats considered in the study. Moreover, Tooth expressions were perceived as more understandable than DNFs in the inspection task (suggesting that they are better suited to tasks that benefit from a more compact representation of knowledge). The same was not generally observed in the classification task. Whilst the time of response was much lower for Tooth expressions than DNFs and the percentage of correct responses was almost the same for Tooth expressions and DNFs, the confidence in the reply and the perceived understandability were higher in the case of DNF formulas. By distinguishing different DNF formats, we observed that longer DNFs (e.g., DNF3) were perceived as less understandable than Tooth expressions. This result was also affected by the background of the respondents. Tooth operators, in particular, resulted in better performances and in a higher level of perceived understandability for users who were not familiar with logic.

The results obtained open several directions for future work. Firstly, we plan a second user study, where both Tooth expressions and DNFs are translated into natural language. Secondly, we plan to compare decision trees and Tooth expressions [15].

References

[1] R. Confalonieri, T. Weyde, T. R. Besold, F. Moscoso del Prado Martín, Using ontologies to enhance human understandability of global post-hoc explanations of black-box models, Artificial Intelligence 296 (2021). doi:https://doi.org/10.1016/j.artint.2021.103471.

[2] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, 2017.

[3] A. Darwiche, P. Marquis, A knowledge compilation map, J. Artif. Intell. Res. 17 (2002) 229–264. doi:10.1613/jair.989.

[4] S. Booth, C. Muise, J. Shah, Evaluating the interpretability of the knowledge compilation map, in: S. Kraus (Ed.), Proc. of IJCAI, 2019, pp. 5801–5807.

[5] D. Porello, O. Kutz, G. Righetti, N. Troquard, P. Galliani, C. Masolo, A toothful of concepts: Towards a theory of weighted concept combination, in: M. Simkus, G. E. Weddell (Eds.), Proceedings of the 32nd International Workshop on Description Logics, Oslo, Norway, June 18-21, 2019, volume 2373 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019.

[6] P. Galliani, O. Kutz, D. Porello, G. Righetti, N. Troquard, On knowledge dependence in weighted description logic, in: D. Calvanese, L. Iocchi (Eds.), GCAI 2019. Proceedings of the 5th Global Conference on Artificial Intelligence, Bozen/Bolzano, Italy, 17-19 September 2019, volume 65 of *EPiC Series in Computing*, EasyChair, 2019, pp. 68–80.

[7] P. Galliani, G. Righetti, O. Kutz, D. Porello, N. Troquard, Perceptron connectives in knowledge representation, in: C. M. Keet, M. Dumontier (Eds.), Knowledge Engineering and Knowledge Management - 22nd International Conference, EKAW 2020, Bolzano, Italy, September 16-20, 2020, Proceedings, volume 12387 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 183–193. doi:10.1007/978-3-030-61244-3_13.

[8] E. Rosch, B. B. Lloyd, Cognition and categorization (1978).

[9] G. Righetti, D. Porello, O. Kutz, N. Troquard, C. Masolo, Pink panthers and toothless tigers: three problems in classification, in: A. Cangelosi, A. Lieto (Eds.), Proc. of the 7th International Workshop on Artificial Intelligence and Cognition, volume 2483 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 39–53.

[10] G. Righetti, C. Masolo, N. Troquard, O. Kutz, D. Porello, Concept combination in weighted logic, in: E. M. Sanfilippo, al. (Eds.), Proc. of the Joint Ontology Workshops 2021 Episode VII, volume 2969 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021.

[11] R. Confalonieri, T. Weyde, T. R. Besold, F. M. del Prado Martín, Trepan Reloaded: A Knowledge-driven Approach to Explaining Black-box Models, in: Proceedings of the 24th European Conference on Artificial Intelligence, 2020, pp. 2457–2464. doi:10.3233/FAIA200378.

[12] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, B. Baesens, An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models, Decis. Support Syst. 51 (2011) 141–154.

[13] H. Allahyari, N. Lavesson, User-oriented assessment of classification model understandability, in: SCAI 2011 Proc., volume 227, IOS Press, 2011, pp. 11–19.

[14] P. Galliani, O. Kutz, N. Troquard, Perceptron operators that count, in: M. Homola, V. Ryzhikov, R. A. Schmidt (Eds.), Proceedings of the 34th International Workshop on Description Logics (DL 2021) part of Bratislava Knowledge September (BAKS 2021), Bratislava,

Slovakia, September 19th to 22nd, 2021, volume 2954 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021.

[15] R. Confalonieri, P. Galliani, O. Kutz, D. Porello, G. Righetti, N. Troquard, Towards knowledge-driven distillation and explanation of black-box models, in: R. Confalonieri, O. Kutz, D. Calvanese (Eds.), Proceedings of the Workshop on Data meets Applied Ontologies in Explainable AI (DAO-XAI 2021) part of Bratislava Knowledge September (BAKS 2021), Bratislava, Slovakia, September 18th to 19th, 2021, volume 2998 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021.