Interpretability of Machine Learning Models and Representations: an Introduction

Conference Paper · April 2016

CITATIONS
92

READS
3,671

2 authors:

Adrien Bibal
University of Colorado
27 PUBLICATIONS **273** CITATIONS
SEE PROFILE

Benoît Frénay
Université Catholique de Louvain - UCLouvain
58 PUBLICATIONS **2,134** CITATIONS
SEE PROFILE

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Interpretability of Machine Learning Models and Representations: an Introduction

Adrien Bibal and Benoît Frénay

Université de Namur - Faculté d'informatique
Rue Grandgagnage 21, 5000 Namur - Belgium

Abstract.
Interpretability is often a major concern in machine learning. Although many authors agree with this statement, interpretability is often tackled with intuitive arguments, distinct (yet related) terms and heuristic quantifications. This short survey aims to clarify the concepts related to interpretability and emphasises the distinction between interpreting models and representations, as well as heuristic-based and user-based approaches.

1 Introduction

According to the literature, measuring the interpretability of machine learning models is often necessary [1], despite the subjective nature of interpretability making such measure difficult to define [2]. Several arguments have been made to highlight the need to consider interpretability alongside accuracy. Some authors note the importance to consider other metrics than accuracy when two models exhibit a similar accuracy [3, 4]. Other authors point out the link between interpretability and the usability of models [5–7]. Often, the medical domain is taken as example. To accept a predictive model, medical experts have to understand the intelligence behind the diagnostic [8], in particular when the decisions surprise them [9]. Furthermore, the detection by experts of anomalies in the model is only possible with interpretable models [10]. Moreover, in some countries, credit denial legally has to be supported by clear reasons, which means that the model supporting this denial has to be interpretable [8]. Finally, it can also be argued that the model itself is a source of knowledge [6, 11, 12].

This survey addresses two issues in the machine learning literature. First, many terms are associated to interpretability, sometimes implicitly referring to different issues. Second, the literature, scattered because of the difficulty to measure interpretability, is neither united nor structured. Although interpretability is often associated with the size of the model, Pazzani wrote in 2000 that "there has been no study that shows that people find smaller models more comprehensibible or that the size of a model is the only factor that affects its comprehensibility" [4]. In 2011, the situation has not changed, according to Huysmans et. al [12] who echo Freitas [11]. Therefore, this survey addresses the two above issues by proposing an unifying and structured view of interpretability focused on models and representations, and concludes by exposing gaps in the literature.

This survey tackles the questions "what is interpretability?" and "how to measure it?". We do not review techniques to make models more interpretable, because we consider the measure of interpretability as being anterior to this

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

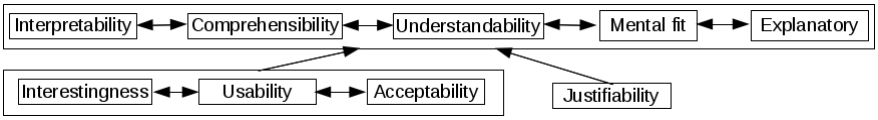

Figure 1: Structure of the main terms used in the literature. $A \rightarrow B$ means that the measure of B requires the measure of A. $A \longleftrightarrow B$ means that measuring A is equivalent to measuring B. Boxes highlight equivalence classes of problems.

problem. In order to answer "what is interpretability?", one needs to gather and unify terms dealing with the problem of interpretability. Section 2 presents several terms used to refer to "interpretability". The second question can be rephrased in terms of comparisons. Sections 3 and 4 review interpretability based on comparisons of models and representations, respectively. Section 5 concludes by highlighting gaps in the literature and corresponding research questions.

2 Different Terms for Different Problems?

This section proposes a unified and structured view of the main terms related to interpretability in the literature. To help researchers when reading papers with distinct terms actually referring to the same problems, a two-level structure is proposed in Fig. 1: the first level consists of synonyms of interpretability and the second level contains terms that rely on interpretability to be measured.

Because of the subjective nature of interpretability, there is no consensus around its definition, nor its measure. First of all, as noted by Rüping [2], the interpretability of a model is not linked to the understandability of the learning process generating this model. Rather, interpretability can be associated to three sub-problems: accuracy, understandability and efficiency [2]. Understandability is central to the problem: an interpretable model is a model that can be understood. Rüping adds accuracy as a necessary criterion in the evaluation of interpretability because "it is always possible to generate a trivial, easily understandable hypothesis without any connection to the data" [2]. Finally, efficiency concerns the time available to the user to grasp the model. Without this criterion, it could be argued that any model could be understood given an infinite amount of time. Other authors use the term interpretability as strict synonym of understandability [5, 10] or comprehensibility [3, 6, 8, 13].

Feng and Michie [14], as other authors after them [15,16], add "mental fit" to the terms interpretability and comprehensibility. Whereas "data fit" corresponds to predictive accuracy [15], "mental fit" is linked to the ability for a human to grasp and evaluate the model [14]. These authors often link interpretability to explainability, e.g. in [15]. An explanatory model "relates attributes to outcomes in a clear, informative, and meaningful way" [17]. According to Ustun and Rudin, interpretability is intuitive for the expert and closely linked to transparency, sparsity, and explanatory [17].

Some other terms are used in combination with interpretability, but actually refer to other problems. Among them, we can consider usability, acceptability

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

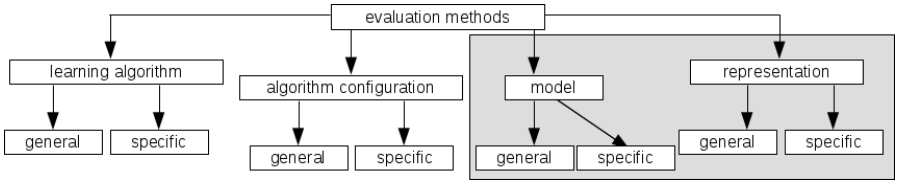

Figure 2: Taxonomy adapted from [20] augmented by representations. Interpretability can be measured for both models and representations (shaded area).

and interestingness. Freitas provides an example where simplicity, often closely linked to interpretability, does not correspond to acceptability for an expert [9]. According to him, following the medical example of [18], experts can be opposed to over-simplistic models. For instance, a three-node tree is probably interpretable, but could be rejected by experts because of its over-simplistic structure [9]. One should note that these two concepts, interpretability and acceptability, are strongly linked but not synonyms, as an acceptable model has to be interpretable, but not vice-versa. In the same way, a model can be considered as not interesting, although being interpretable.

Finally, the term "justifiability" can also be observed alongside interpretability, as it requires an expert to assess that the model "is in line with existing domain knowledge" [8, 19]. As for usability and interestingness cited above, justifiability depends on the interpretability of the model [8].

3 Comparing Models in Terms of Interpretability

Even if we agree on terms to discuss interpretability, one still needs an actual measure of interpretability. In general, measures can be applied on many components of learning systems. Lavesson and Davidsson propose a taxonomy for evaluation methods that classifies them according on whether they assess learning algorithms, algorithm configurations (meta-parameters) or models [20]. Those three elements can be either specific or general, like e.g. evaluation methods for a specific type of model or for distinct types. Similarly, we believe that it is necessary to make a clear distinction between interpretability measures depending on what specific component they target. In this view, we extend the taxonomy of Lavesson and Davidsson by also considering representations in Fig. 2 inspired by [20]: measures of interpretability can be applied to either models or representations through specific approaches. First, comparing mathematical entities such as models requires to define quantitative measurements. This approach is one of the two approaches highlighted by Freitas [9] and can be called the heuristic approach [2]. The second approach uses user-based surveys to assess the interpretability of models. However, unlike the first approach, the models are evaluated through their representations. This second approach is closely linked to information visualisation. This section considers interpretability of models and Section 4 deals with interpretability of their representations.

The heuristic approach can compare models from the same type, e.g. two

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

SVM models. The size of the model is one of the most used heuristic [2, 6]. For instance, two decision rule lists/sets can be compared in terms of their number of rules and terms [21, 22] and two decision trees can be compared in terms of their number of nodes [23]. Some authors base their heuristics on the psychological theory of Miller, stating that human beings can only deal with 7 ± 2 abstract entities at the same time [24]. For instance, Wheis and Sondhauss propose a maximum of 7 in the number of dimensions [15]. Another way to evaluate the complexity of models is the minimum description length (MDL) [20], but the result depends on the coding scheme for the model parameters, also making this technique specific to the model type [20].

Comparing models of distinct types is more challenging, as the characteristics related to the interpretability of a model from a certain type can be missing in the model from another type. For instance, one cannot minimise the number of nodes of a SVM model. To overcome this difficulty, Backhaus and Seiffert propose to consider three generic criteria: "the ability of the model to select features from the input pattern, the ability to provide class-typical data points and information about the decision boundary directly encoded in model parameters" [25]. For instance, SVM models are graded 1 out of 3, because they only satisfy the third criterion thanks to the "stored support vectors and kernel" [25]. SVM models compete with other models ranked 1 out of 3, but are less interpretable than others ranked 2 or 3 out of 3. Whereas this ranking is able to compare models of distinct types, it does not allow to compare the interpretability of models from the same type. Other limitations of heuristics exist, i.e. they deal with "syntactical interpretability" and do not consider semantic interpretability [9].

4 Comparing Representations in Terms of Interpretability

The limitations depicted in Section 3 are overcome by a measure based on users that evaluate models through their representations. Allahyari and Lavesson used a survey filled by users to evaluate the interpretability of models generated by 6 learning algorithms [26]. This user-based study compared models pairwise by asking questions like "is this model more understandable than the other one?" [26]. Such surveys allow comparing models of the same type, but also models of distinct types. Following the same idea, Piltaver et. al. designed a survey [27] validated [13] on decision trees. Huysmans et. al. checked the accuracy, answer time and answer confidence of users who were asked to grasp a certain model through its representation [12]. Other questions evaluate the understanding of the model. These authors compared the interpretability of three different representations: trees, decision tables and textual representation of rules. Thanks to this kind of evaluation, they could check the link between interpretability and the simplicity of the model, and could conclude by asking questions such as: "to what extent the representations discussed in this study continue to remain useful once they exceed a certain size" [12]. This new trend echoes Rüping when he wrote that "due to the informal nature of the concept of interpretability, a survey over human expert is the most promising measure" [2].

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

Evaluating representations before evaluating accuracy of models has a particular advantage. Following the idea of Wheis and Sondhauss [15], one could first choose the type of representation having the highest interpretability for a certain group of users, and only then select the type of model having the highest accuracy among those that can be represented by the selected representation.

As a final remark, one could argue that, in the context of interpretability, there is no such thing as a comparison of models, but only comparisons of representations. One can go further and only compare visualisations of model representations, since representations can be either uninterpretable or highly interpretable depending on the way they are shown to the user. Yet, it should be noted that the user-based approach (comparing representations and visualisations thereof) does not allow to quantify interpretability. In contrast, heuristics can be integrated in learning through multi-objective optimisation techniques.

5 Conclusion

This paper presents two major difficulties in the measure of interpretability. First, distinct terms are used in the literature. We separated them into the ones used as strict synonyms (e.g. understandability and comprehensibility) and the ones that depend on interpretability to be defined but related to distinct problems (e.g. justifiability and usability). Second, papers in the literature can be divided into comparisons of the interpretability of models and representations, that is comparisons based on mathematical heuristics or user-based surveys.

In the literature, there is no clear-cut distinction between the interpretability measure of models and representations. The two research questions "what is an interpretable model?" and "what is an interpretable representation?" need to be investigated independently. Furthermore, many papers rely on intuition in the use of interpretability, which leads to a focus on "white-boxes" (decision trees, decision rules, etc.) and a lack of consideration of "black-boxes" (SVM, neural networks, etc.). This distinction would benefit from a grey-scale approach. Finally, there is a lack of literature around user-based measures of interpretability, leaving the question "do heuristics accurately model the understanding of users?" with almost no answer. There is a need to link the results of user-based surveys with heuristics in order to translate the former into the latter and hopefully optimise mathematically the interpretability described by users.

References

[1] Y. Kodratoff. The comprehensibility manifesto. *AI Communications*, 7(2):83–85, 1994.

[2] S. Rüping. *Learning interpretable models*. PhD thesis, Universität Dortmund, 2006.

[3] C. Giraud-Carrier. Beyond predictive accuracy: what? In *Proc. ECML*, pages 78–85, Chemnitz, Germany.

[4] M. J. Pazzani. Knowledge discovery from data? *IEEE Intelligent Systems and their Applications*, 15(2):10–12, 2000.

[5] G. Nakhaeizadeh and A. Schnabl. Development of multi-criteria metrics for evaluation of data mining algorithms. In *Proc. KDD*, pages 37–42, Newport Beach, CA, USA, 1997.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

[6] I. Askira-Gelman. Knowledge discovery: comprehensibility of the results. In *Proc. HICSS*, volume 5, pages 247–255, Maui, HI, USA, 1998.

[7] A. Vellido, J. D. Martin-Guerroro, and P Lisboa. Making machine learning models interpretable. In *Proc. ESANN*, pages 163–172, Bruges, Belgium, 2012.

[8] D. Martens, J. Vanthienen, W. Verbeke, and B. Baesens. Performance of classification models from a user perspective. *Decision Support Systems*, 51(4):782–793, 2011.

[9] A. A. Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD Explorations Newsletter*, 15(1):1–10, 2014.

[10] A. Andrzejak, F. Langner, and S. Zabala. Interpretable models from distributed data via merging of decision trees. In *Proc. CIDM*, pages 1–9, Singapore, 2013.

[11] A. A Freitas. Are we really discovering interesting knowledge from data? *Expert Update*, 9(1):41–47, 2006.

[12] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141–154, 2011.

[13] R. Piltaver, M. Luštrek, M. Gams, and S. Martinčić-Ipšić. Comprehensibility of classification trees - survey design validation. In *Proc. ITIS*, pages 5–7, Šmarješke toplice, Slovenia, 2014.

[14] C. Feng and D. Michie. Machine learning of rules and trees. *Machine Learning, Neural and Statistical Classification. Ellis Horwood, Hemel Hempstead*, 1994.

[15] C. Weihs and U.M. Sondhauss. Combining mental fit and data fit for classification rule selection. In *Exploratory Data Analysis in Empirical Research*, pages 188–203. Springer, 2003.

[16] O. Maimon and L. Rokach. Decomposition methodology for knowledge discovery and data mining. In *Data Mining and Knowledge Discovery Handbook*, pages 981–1003. Springer, 2005.

[17] B. Ustun and C. Rudin. Methods and models for interpretable linear classification. *arXiv preprint arXiv:1405.4047*, 2014.

[18] T. Elomaa. In defense of c4. 5: Notes on learning one-level decision trees. In *Proc. ICML*, pages 62–69, Beijing, China, 2014.

[19] D. Martens, M. De Backer, R. Haesen, B. Baesens, C. Mues, and J. Vanthienen. Ant-based approach to the knowledge fusion problem. In *Proc. ANTS*, pages 84–95, Brussels, Belgium, 2006.

[20] N. Lavesson and P. Davidsson. Evaluating learning algorithms and classifiers. *International Journal of Intelligent Information and Database Systems*, 1(1):37–52, 2007.

[21] M. Schwabacher and P. Langley. Discovering communicable scientific knowledge from spatio-temporal data. In *Proc. ICML*, pages 489–496, Williamstown, MA, USA, 2001.

[22] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan. An interpretable stroke prediction model using rules and bayesian analysis. In *Proc. AAAI*, Bellevue, WA, USA, 2013.

[23] A. Van Assche and H. Blockeel. Seeing the forest through the trees: Learning a comprehensible model from an ensemble. In *Proc. ECML*, pages 418–429, Warsaw, Poland, 2007.

[24] G. A Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81–97, 1956.

[25] A. Backhaus and U. Seiffert. Classification in high-dimensional spectral data: Accuracy vs. interpretability vs. model size. *Neurocomputing*, 131:15–22, 2014.

[26] H. Allahyari and N. Lavesson. User-oriented assessment of classification model understandability. In *Proc. SCAI*, pages 11–19, Trondheim, Norway, 2011.

[27] R. Piltaver, M. Luštrek, M. Gams, and S. Martinčić-Ipšić. Comprehensibility of classification trees–survey design. In *Proc. IS*, pages 70–73, Ljubljana, Slovenia, 2014.