Minimal Rules from Decision Forests: a Systematic Approach

Giovanni Pagliarini[1], Andrea Paradiso[1], Marco Perrotta[1] and Guido Sciavicco[1]

[1]University of Ferrara, Italy

Abstract

Extracting logical rules from ensembles of symbolic learning models, and especially from ensembles of decision trees, is a very well-known discipline, and several methods and algorithms have been proposed for its solution. However, the existing approaches are characterized by being purely statistical. In this paper, we discuss the problem of systematically extracting minimal logical rules from ensembles of trees from both a theoretical and an algorithmic point of view.

Keywords

minimal logical rules, ensembles of symbolic models, forests of trees

1. Introduction

In sharp opposition to the proliferation of machine learning models used to approach all range of typical artificial intelligence problems, from classification, to regression, to reinforcement learning, the universal concept linked to the interpretation and the explanation of such models is that of *rule*. As a matter of fact, it can be said that all different statistical learning models are different methods for implicit or explicit rule extraction from data.

Unsurprisingly, the idea of expressing the behaviour of a data-driven artificial intelligent agent in terms of rules is extremely pervasive. On the one side, learning models are usually separated into *symbolic* ones, such as decision trees or linear regressions, *sub-symbolic* ones, such as neural networks, and *mixed* ones, such as ensembles of trees. On the other side, methods for explaining learning models are classified into *global* ones, that are focused on the model as a whole, and *local* ones, focused on the behaviour of a model on a specific instance. Numerous literature surveys on explainable artificial intelligence and interpretable machine learning have been conducted (e.g., see [1, 2, 3]).

With symbolic learning methods we represent data and relationships using symbolic structures. *Decision trees* [4] are a classic example of symbolic learning models, where the tree branches typically represent formulas of propositional logic. Despite their effectiveness, decision trees suffer from a limited ability to generalize to new data. To address such an issue, ensembles of independent decision trees, known as *decision forests*, are commonly used to improve the generalization ability of single trees. Decision forests are symbolic in nature, but they include a functional component to amalgamate the output of single trees, and can be therefore classified as mixed symbolic/sub-symbolic techniques; the most famous algorithm for decision forest learning, namely *random forest* [5], produces decision forests for classification/regression in which the aggregation function is simple majority.

The development of global explanation methods for decision forests is paramount, and several solutions have been proposed in the past, including *Partial Dependence Plots* [6], *Accumulated Local Effects Plots* [7], global surrogate models [8], and *SHapley Additive exPlanations* (*SHAP*) [9]. In terms of rule extraction from decision forests, the relevant methods include the celebrated *Simplified Tree Ensemble Learner (STEL)* [10] (recently extended to the modal case in [11]), and several heuristic

OVERLAY 2024, 6th International Workshop on Artificial Intelligence and Formal Verification, Logic, Automata, and Synthesis, November 28–29, 2024, Bolzano, Italy

giovanni.pagliarini@unife.it (G. Pagliarini); andrea.paradiso@edu.unife.it (A. Paradiso); marco.perrotta@edu.unife.it (M. Perrotta); guido.sciavicco@unife.it (G. Sciavicco)

0000-0002-8403-3250 (G. Pagliarini); 0000-0002-3614-2487 (A. Paradiso); 0009-0009-1497-5291 (M. Perrotta); 0000-0002-9221-879X (G. Sciavicco)

techniques [12, 13]. Two common traits characterize this plethora of proposals: first, global rules from decision forests are extracted in statistical form, and, second, these methods are seldom integrated into widely used machine learning portfolios and frameworks.

We initiate a systematic study of purely logical global rules extraction from decision forests. The language of standard decision forest is propositional; propositions range from simple assertions of the type $A \bowtie a$, where A is an *attribute*, a is a constant, and $\bowtie$ is a comparison operator (e.g., *fever is over 38 degrees*), to complex (yet atomic) sentences, of the type $f(A_1, \ldots, A_k) \bowtie a$, where $A_1, \ldots, A_k$ are attributes and f is an (arbitrary) function applied to them (e.g., *the averaged vibration of sensors A and B is below 100 Hertz* — examples of such propositions emerge, among others, in oblique decision trees and forests [14]). In any case, atomic propositions form a simple theory $\mathcal{T}$ (e.g., *fever is over 38 degrees* implies *fever is over 37 degrees*); this is ignored during the learning phase, which generates a certain amount of redundancy in learnt models. In the simple and representative case of decision forests for classification, that is, learning a class L from a dataset $\mathcal{I}$, given a decision forest F learnt from $\mathcal{I}$ we define a *strong class rule* for it as an object of the type $\varphi \Leftrightarrow L$, where L is a class and φ is a propositional formula, such that F classifies an instance $I \in \mathcal{I}$ as L if and only if I satisfies φ. Obviously, it is to be expected that useful class rules have relatively short antecedent, and since the latter is a propositional formula, it can be *minimized*. Minimization of propositional formulas is a very well-known problem. In the most common case the size of a formula φ, denoted by $|\varphi|$, is defined as the number of its symbols, and the minimization problem asks, given a propositional formula φ: which is a formula φ', equivalent to φ (denoted $\varphi' \equiv \varphi$), and minimal in size? In its decision version, the problem becomes, given φ and a number q: does there exists a formula φ', such that $|\varphi'| \leq q$ and $\varphi' \equiv \varphi$? This problem is Σ_p^2-complete, and there exist a number of approaches for it [15].

In this paper we ask the question, given a decision forest F and a class L: which is a strong class rule for L whose antecedent φ is minimal in size? In its decision version, given a number q, it becomes: is there a strong class rule for L whose antecedent φ is such that $|\varphi| \leq q$? Obtaining small class rules differs from pure logical minimization in two key aspects. First, minimization of propositional formula is usually not intended *modulo* a theory $\mathcal{T}$, but, in general, classic minimization methods and techniques can be adapted to this case. Second, in our case minimization does not need to preserve logical equivalence, but only equivalence modulo the set of instances $\mathcal{I}$ on which the original forest was learnt. We shall see that this problem can be very hard in terms of computational complexity, and it makes sense to consider other (ideally, simpler) versions of it. We define the concept of *right weak class rule*, (resp., *left weak class rule*) that is, a rule of the type $\varphi \Rightarrow L$ (resp., $\varphi \Leftarrow L$) such that if $I \in \mathcal{I}$ satisfies φ (resp., F classifies I as L) then F classifies I as L, (resp., that if $I \in \mathcal{I}$ satisfies φ), and we ask the question, given a decision forest F and a class L: which is a right (resp., left) weak class rule for L whose antecedent φ is minimal (resp., non-trivially maximal) in size? Or, in its decision version, given also a number q: is there a right (resp., left) weak class rule for L whose antecedent (resp., non-trivial antecedent) φ is such that $|\varphi| \leq q$ (resp., $|\varphi| \geq q$)?

2. Decision Trees, Decision Forests, and Class Formulas

Definition 1. *A* dataset *is a set of* m instances $\mathcal{I} = \{I_1, \ldots, I_m\}$*, each one of which is described by the values of* n attributes $\mathcal{A} = \{A_1, \ldots, A_n\}$.

Without lack of generality we assume that the value of each attribute in an instance is a real number. Several problems are usually associated with datasets; in the case of *supervised* learning, each instance is also associated to a *label* (or *class*) $L \in \mathcal{L}$, and a dataset is termed *labelled*. Given a labelled dataset $\mathcal{I}$, supervised *classification* consists of synthesizing an algorithm (a *classifier*) that is able to classify the instances of an unlabelled dataset $\mathcal{J}$ whose instances are defined on the same set of attributes.

In the symbolic context, instances are seen as logical models. To help this interpretation one takes into consideration that datasets are naturally associated to a logical vocabulary $\mathcal{P}$ of propositional

letters, from which formulas are built. In the most general case, we have

$$\mathcal{P} \quad = \quad \{(f(A_1, \ldots, A_k) \bowtie \ a) \mid f \in \mathcal{F}, a \in \mathbb{R}, \bowtie \in \{<, \leq, =, \geq, >\}\},$$

where $\mathcal{F}$ is a set of suitable *feature extraction functions*. To a dataset $\mathcal{I}$, we associate its vocabulary $\mathcal{P}$ and a (possibly empty) *theory* $\mathcal{T}$, that is, a set of propositional formulas of the type $p_i \rightarrow p_j$, where $p_i, p_j \in \mathcal{P}$, that expresses semantic constraints between propositional letters (e.g., $A > 5$ implies $A > 4$). In the following, we write $I \models \varphi$ to denote that a propositional formula φ is satisfied by I.

Definition 2. *Let $\mathcal{L}$ be a set of classes and $\mathcal{P}$ a finite set of propositional letters. Then, a* decision tree *(*on *$\mathcal{L}$) is a tuple*

$$\tau = \langle V, E, l, e \rangle,$$

where $\langle V, E \rangle$ is a full binary directed tree, l is a leaf-labelling function *that assigns a class from $\mathcal{L}$ to each leaf node in V, and e is an* edge-labelling function *that assigns a* decision *from $\{p, \neg p \mid p \in \mathcal{P}\}$ to each edge in E, in such a way that two siblings always have opposite decisions. A* decision forest *$F = \{\tau_1, \ldots, \tau_z\}$ (*on *$\mathcal{L}$) is a set of z decision trees (*on *$\mathcal{L}$).*

A decision tree/forest is learnt from a dataset $\mathcal{I}$ based on its vocabulary $\mathcal{P}$ (the *language of* the tree/forest). An instance is classified by a tree by progressively checking the truth value of each proposition on a path (and we denote with $\tau(I)$ the class assigned to an instance I by τ), and by a forest ($F(I)$) by systematically querying each tree individually and then aggregating their decisions; among other possibilities, a typical aggregation function is simple majority, which is assumed here.

Definition 3. *Given a decision tree τ on $\mathcal{L}$ and a class $L \in \mathcal{L}$, then: (i) given and a path π in τ from the root to a leaf labeled with L (an* L-path*), the conjunction of all decisions on π is called* L-path tree formula*, and it is denoted by φ^L_π, and (ii) the disjunction of all L-path formulas in τ is called* L-class tree formula *(φ^L_τ). Given a decision forest F on $\mathcal{L}$ with z trees, and a class $L \in \mathcal{L}$, then: (i) given a collection $\tau_{i_1}, \ldots, \tau_{i_t} \in F$ and one L-path π_{i_j} ($1 \leq j \leq t$) per tree τ_{i_j}, the conjunction of all L-path tree formulas $\varphi^L_{\pi_{i_1}}, \ldots, \varphi^L_{\pi_{i_1}}$ is called* partial L-path forest formula *($\varphi^L_{\pi_{i_1}, \ldots, \pi_{i_t}}$); (ii) if $t > z/2$, then $\varphi^L_{\pi_{i_1}, \ldots, \pi_{i_t}}$ is called* L-path forest formula*; and (iii) the disjunction of all possible L-path forest formulas is called* L-class forest formula *(φ^L_F).*

3. Rules from Decision Forests

Definition 4. *Given a decision forest F, learnt from a dataset $\mathcal{I}$, on $\mathcal{L}$, and a class $L \in \mathcal{L}$, a* strong L-class rule *is an object of the type $\varphi \Leftrightarrow L$, where φ (the* antecedent*) is a propositional formula in the language of F, such that, for every $I \in \mathcal{I}$, $I \models \varphi$ if and only if $F(I) = L$, a* right weak L-class rule *is an object of the type $\varphi \Rightarrow L$ such that $I \models \varphi$ implies $F(I) = L$, and a* left weak L-class rule *is an object of the type $L \Rightarrow \varphi$ such that $F(I) = L$ implies $I \models \varphi$.*

Given a decision forest F on $\mathcal{L}$, a class $L \in \mathcal{L}$, and the L-class forest formula φ^L_F, $\varphi^L_F \Leftrightarrow L$ is a (trivial) strong L-class rule. In practical terms, it will likely be very redundant, due to the fact that decision trees and forest are general learnt via sub-optimal learning algorithms (recall that the problem of extracting a minimal decision tree is NP-hard, and sub-optimal, polynomial algorithms are commonly used for learning), and the fact that the theory $\mathcal{T}$ underlying the language is ignored during learning. Thus, given a decision forest F on $\mathcal{L}$, learnt from a dataset $\mathcal{I}$, and a class $L \in \mathcal{L}$, we are interested in finding a *minimal* strong L-class rule, that is, a strong L-class rule $\varphi \Leftrightarrow L$ such that, for every strong class rule $\varphi' \Leftrightarrow L$ so that $\varphi \equiv^{\mathcal{I}}_{\mathcal{T}} \varphi'$ (i.e., so that φ and φ' are equivalent modulo $\mathcal{T}$ at least with respect to the instances in $\mathcal{I}$), it is the case that $|\varphi| \leq |\varphi'|$.

One way to assess the complexity of the problem of finding minimal strong rules is to study its decision version, that is: given a decision forest F, learnt from a given dataset $\mathcal{I}$, on $\mathcal{L}$, a class $L \in \mathcal{L}$, and a number q, is there a strong L-class rule $\varphi \Leftrightarrow L$ such that $|\varphi| \leq q$? Given that the size of the input of this problem is the number of symbols of F (denoted by $|F|$) plus the number of instances in $\mathcal{I}$ ($|\mathcal{I}|$) and the size of the representation of q ($|q|$), we have the following result.

Theorem 1. *Given a decision forest F, learnt from a dataset $\mathcal{I}$, on $\mathcal{L}$, a class $L \in \mathcal{L}$, a theory $\mathcal{T}$, and a number q, the problem of establishing if there exists a strong L-class rule $\varphi \Leftrightarrow L$ such that $|\varphi| \leq q$ is in NEXPTIME.*

Proof[sketch]. An (DNF) antecedent φ of size less than or equal to q can be guessed. Then, for each instance $I \in \mathcal{I}$, I is checked against both F and φ: if $F(I) = L$ (resp., not L) and $I \models \varphi$ (resp., $I \not\models \varphi$), then I is marked. If all instances in $\mathcal{I}$ end up being marked, then $\varphi \Rightarrow F$ is a strong L-class rule. The complexity of this process is polynomial in $|F|$ and $|\mathcal{I}|$, but exponential in $|q|$, and the value of the minimal q for which a strong L-class rule exist may be exponential in $|F|$. □

Since extracting a minimal strong class rule may turn out to be impractical, we turn our attention to weak rules. Given a decision forest F, learnt on a dataset $\mathcal{I}$, on $\mathcal{L}$, a class $L \in \mathcal{L}$, and a L-path forest formula $\varphi^L_{\pi_{i_1},\ldots,\pi_{i_t}}$, $\varphi^L_{\pi_{i_1},\ldots,\pi_{i_t}} \Rightarrow L$ is a (trivial) right weak L-class rule whose antecedent may exhibit the same kind of redundancy and can be minimized as in the previous case. Similarly, $\top \Leftarrow L$ is a (trivial) left weak L-class rule whose antecedent can be maximized; in this case, however, one is interested in non-trivial maximal antecedents.

Theorem 2. *Given a decision forest F on $\mathcal{L}$, a class $L \in \mathcal{L}$, a theory $\mathcal{T}$, and a number q, the problem of establishing if there exists a right weak L-class rule $\varphi \Rightarrow L$ such that $|\varphi| \leq q$ is in NP, and the problem of establishing if there exists a non-trivial left weak L-class rule $L \Rightarrow \varphi$ such that $|\varphi| \geq q$ is in NEXPTIME.*

Proof[sketch]. As for right weak L-class rules, a (term) antecedent φ of size less than or equal to q can be guessed. Then, for each instance $I \in \mathcal{I}$, I is checked against both F and φ: if $I \not\models \varphi$, or $I \models \varphi$ and $F(I) = L$, then I is marked. If all instances in $\mathcal{I}$ end up being marked, then $\varphi \Rightarrow F$ is a weak L-class rule. The complexity of this process is polynomial in $|F|$, $|\mathcal{I}|$, but exponential in $|q|$; however, the value of the minimal q for which a right weak L-class rule exist is polynomial in $|F|$. As for left weak L-class rules, a (DNF) antecedent φ of size grater than or equal to q can be guessed. Then, we first check that φ is non-trivial, that is, there are no repeated literals in any term, no term is unsatisfiable, and no terms implies any other term. Then, for each instance $I \in \mathcal{I}$, I is checked against both F and φ: if $F(I) \neq L$, or $F(I) = L$ and $I \models \varphi$, then I is marked. If all instances in $\mathcal{I}$ end up being marked, then $\varphi \Rightarrow F$ is a weak L-class rule. The complexity of this process is polynomial in $|F|$ and $|\mathcal{I}|$, but exponential in $|q|$, and the value of the minimal q for which a left weak L-class rule exist may be exponential in $|F|$. □

4. Conclusions

We started a systematic study of logical methods for rule extraction from decision forests, a well-known classification model. Extracting rules from decision forests is a well-known problem, but existing solutions are statistical and data-driven. In our work, we apply known logical algorithms to rule extraction, contributing to bridging the gap between logic and machine learning.

Acknowledgments

We acknowledge the support of the FIRD project *Methodological Developments in Modal Symbolic Geometric Learning*, funded by the University of Ferrara, and the INDAM-GNCS project *Symbolic and Numerical Analysis of Cyberphysical Systems* (code CUP_E53C23001670001), funded by INDAM; Giovanni Pagliarini and Guido Sciavicco are GNCS-INdAM members. Moreover, this research has also been funded by the Italian Ministry of University and Research through PNRR - M4C2 - Investimento 1.3 (Decreto Direttoriale MUR n. 341 del 15/03/2022), Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 8 "Pervasive AI", funded by the European Union under the NextGeneration EU programme".

References

[1] D. V. Carvalho, E. M. Pereira, J. S. Cardoso, Machine Learning Interpretability: A Survey on Methods and Metrics, Electronics 8 (2019) 1–34.

[2] M. Du, N. Liu, X. Hu, Techniques for interpretable machine learning, Communications of the ACM 63 (2020) 68–77.

[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A Survey of Methods for Explaining Black Box Models, ACM Computing Surveys 51 (2019) 93:1–93:42.

[4] J. Quinlan, Induction of Decision Trees, Machine Learning 1 (1986) 81–106.

[5] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.

[6] J. Friedman, Greedy function approximation: A gradient boosting machine., The Annals of Statistics 29 (2001) 1189–1232.

[7] D. Apley, J. Zhu, Visualizing the effects of predictor variables in black box supervised learning models, Journal of the Royal Statistical Society Series B 82 (2020) 1059–1086.

[8] M. Craven, J. Shavlik, Extracting Tree-Structured Representations of Trained Networks, in: Proceedings of the 8th Advances in Neural Information Processing Systems (NIPS), 1995, pp. 24–30.

[9] S. Lundberg, S. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), 2017, pp. 4768–4777.

[10] H. Deng, Interpreting tree ensembles with inTrees, International Journal of Data Science and Analytics 7 (2019) 277–287.

[11] M. Ghiotti, F. Manzella, G. Pagliarini, G. Sciavicco, I. Stan, Evolutionary explainable rule extraction from (modal) random forests, in: Proc. of the 26th European Conference on Artificial Intelligence (ECAI) and the 12th Conference on Prestigious Applications of Intelligent Systems (PAIS), volume 372 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2023, pp. 827–834.

[12] M. Mashayekhi, R. Gras, Rule extraction from decision trees ensembles: New algorithms based on heuristic search and sparse group lasso methods, International Journal of Information Technology & Decision Making 16 (2017) 1707–1727.

[13] C. Bénard, G. Biau, S. Da Veiga, E. Scornet, Interpretable random forests via rule extraction, in: Proc. of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS), volume 130 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 937–945.

[14] B. Menze, B. Kelm, D. Splitthoff, U. Koethe, F. Hamprecht, On oblique random forests, in: Proc. of the European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2011, pp. 453–469.

[15] C. Umans, T. Villa, A. Sangiovanni-Vincentelli, Complexity of two-level logic minimization, IEEE Transactions on Compututer Aided Design of Integrated Circuits Systems 25 (2006) 1230–1246.