

On Trustworthy Rule-Based Models and Explanations

Mohamed Siala, Jordi Planes, and Joao Marques-Silva

September 18, 2025

LAAS CNRS & INSA Toulouse
University of Lleida
ICREA

Explainability is Not a Game

When the decisions of ML models impact people, one should expect explanations to offer the strongest guarantees of rigor. However, the most popular XAI approaches offer none.

Fairwashing: the risk of rationalization

Ulrich Aivodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, Alain Tapp

Proceedings of the 36th International Conference on Machine Learning, PMLR 97:161-170, 2019.

Against Interpretability: a Critical Examination of the Interpretability Problem in Machine Learning

Research Article | [Open access](#) | Published: 13 August 2019

Volume 33, pages 487–502, (2020) [Cite this article](#)

The (Un)reliability of Saliency Methods

Chapter | First Online: 10 September 2019

pp 267–280 | [Cite this chapter](#)

Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods

Authors: [Dylan Slack](#), [Sophie Hilgard](#), [Emily Jia](#), [Sameer Singh](#), [Himabindu Lakkaraju](#) [Authors Info & Claims](#)

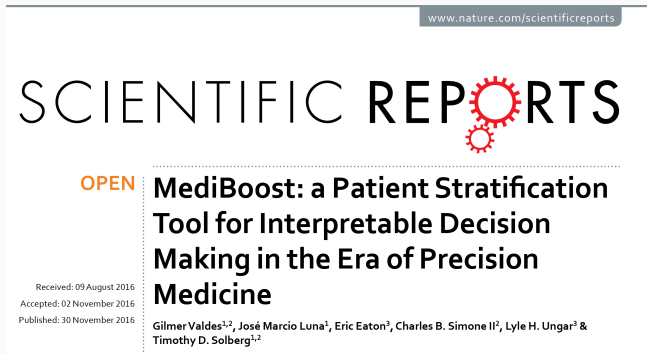
AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society • Pages 180 - 186
<https://doi.org/10.1145/3375627.3375830>

An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models

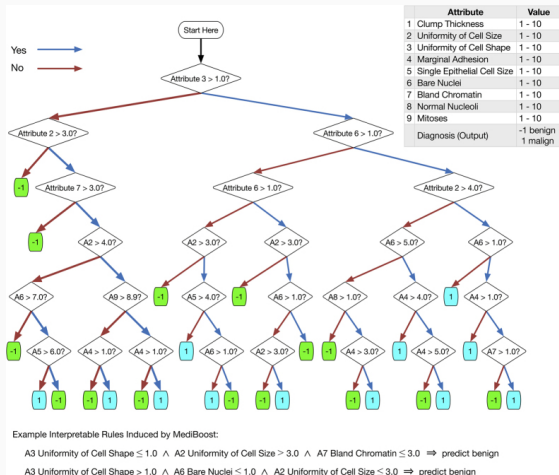
Authors: [Johan Huysmans](#), [Karel Dejaeger](#), [Christophe Mues](#), [Jan Vanthienen](#), [Bart Baesens](#) [Authors Info & Claims](#)

Decision Support Systems, Volume 51, Issue 1 • Pages 141 - 154 • <https://doi.org/10.1016/j.dss.2010.12.003>

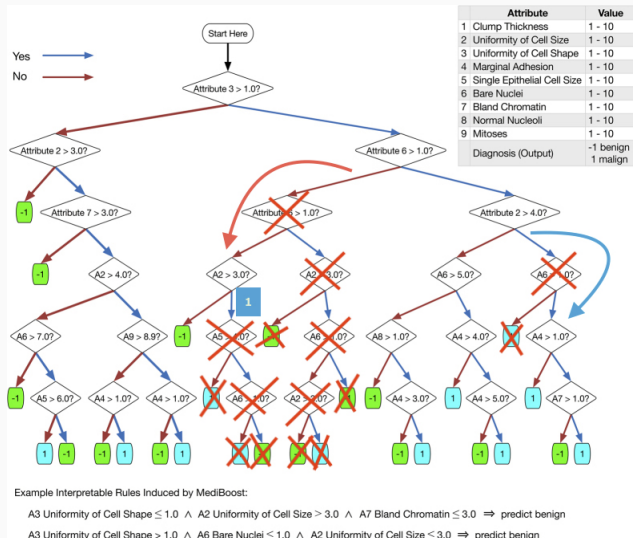
Two Facets of Trust in Rule-based Models and Explanations



Redundancy



Redundancy



Overlap

- **Adult dataset:** The task is to predict whether an individual's annual income exceeds \$50,000.

- **Adult dataset:** The task is to predict whether an individual's annual income exceeds \$50,000.
- **Illustrative examples of Anchor explanations**, produced by a neural network trained on the Adult dataset:
 - $Age \leq 28 \wedge Occupation = Other \wedge CapitalGain = 0 \wedge CapitalLoss = 0 \wedge Workclass = Private \implies 0$
 - $MaritalStatus = Married-civ-spouse \wedge Education = Masters \implies 1$

- **Adult dataset:** The task is to predict whether an individual's annual income exceeds \$50,000.
- **Illustrative examples of Anchor explanations**, produced by a neural network trained on the Adult dataset:
 - $Age \leq 28 \wedge Occupation = Other \wedge CapitalGain = 0 \wedge CapitalLoss = 0 \wedge Workclass = Private \implies 0$
 - $MaritalStatus = Married-civ-spouse \wedge Education = Masters \implies 1$
- In such cases, the resulting explanations may lack reliability and cannot be considered fully trustworthy.

We answer the following questions:

We answer the following questions:

- Is it possible to develop a general-purpose approach for eliminating redundancy in rule-based models and explanations under background constraints?

We answer the following questions:

- Is it possible to develop a general-purpose approach for eliminating redundancy in rule-based models and explanations under background constraints?
- What is the relationship between rule succinctness and formal explainability?

We answer the following questions:

- Is it possible to develop a general-purpose approach for eliminating redundancy in rule-based models and explanations under background constraints?
- What is the relationship between rule succinctness and formal explainability?
- Can we identify and characterize negative overlaps between rules or explanations in the feature space, subject to background constraints?

We answer the following questions:

- Is it possible to develop a general-purpose approach for eliminating redundancy in rule-based models and explanations under background constraints?
- What is the relationship between rule succinctness and formal explainability?
- Can we identify and characterize negative overlaps between rules or explanations in the feature space, subject to background constraints?
- How can these ideas be effectively implemented in practice?

We answer the following questions:

- Is it possible to develop a general-purpose approach for eliminating redundancy in rule-based models and explanations under background constraints?
- What is the relationship between rule succinctness and formal explainability?
- Can we identify and characterize negative overlaps between rules or explanations in the feature space, subject to background constraints?
- How can these ideas be effectively implemented in practice?
- Do well-known tools suffer from redundancy and overlap, and to what extent?

Propositionnnal Logic

Propositionnnal Logic

- Let F_1 and F_2 be two Boolean formulas. $F_1 \models F_2$ if every solution of F_1 is a solution of F_2

Propositionnnal Logic

- Let F_1 and F_2 be two Boolean formulas. $F_1 \models F_2$ if every solution of F_1 is a solution of F_2
- There are very efficient tools that can be used to check if $F_1 \models F_2$. They are called SAT (Boolean Satisfiability Solvers).

Rule-Based Models

- We consider models that can be represented as a set of (unordered) rules together with background constraints \mathcal{B}
- Examples include decision trees, decision diagrams, random forests, boosted trees, decision sets, among others
- We use \mathcal{M} to denote a model
- A literal is a unary relation on a feature. For instance $(size > 20)$ is a literal.
- A rule R is denoted by $L \implies o$, where L is a conjunction of literals and o is a prediction outcome

Example

Background constraints \mathcal{B}

- $(salary > 0) \leftrightarrow (age \geq 18)$
- $(size = 140) \rightarrow (size > 120)$
- $(weight > 90) \rightarrow (weight \geq 85)$
- $(weight \geq 85) \rightarrow (weight > 80)$

The Model

- $R_1 = (salary > 0) \wedge (size \neq 140) \wedge (age > 10) \wedge (color = blue) \wedge (weight > 80) \implies 1$
- $R_2 = (salary > 0) \wedge (size = 140) \implies 1$
- $R_3 = (salary > 0) \wedge (weight > 90) \implies 1$
- $R_4 = (size > 120) \wedge (weight < 85) \implies 0$

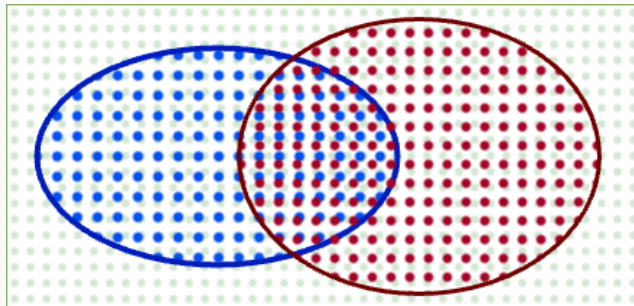
Theoretical Contributions On Overlap and Redundancy

Overlap

- Given two rules R_1 and R_2 , do R_1 and R_2 overlap?
- Can we find all negative overlaps?

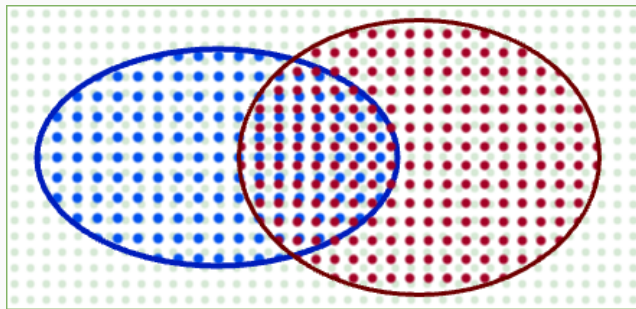
Overlap

Figure 1: Illustration of overlap between two rules R_1 and R_2 . Points are inputs in feature space that satisfy \mathcal{B} . Blue: fire R_1 ; Red: fire R_2 .



Overlap

Figure 1: Illustration of overlap between two rules R_1 and R_2 . Points are inputs in feature space that satisfy \mathcal{B} . Blue: fire R_1 ; Red: fire R_2 .



Lemma (Overlap Check)

Two rules R_1 and R_2 overlap iff $\mathcal{B} \wedge L_1 \wedge L_2$ is satisfiable.

- Rule redundancy
- Literal redundancy
 - Local redundancy
 - Global redundancy

Rule Redundancy

Definition (Rule Redundancy)

A rule R is redundant in \mathcal{M} iff $\mathcal{M} \setminus R$ is equivalent to \mathcal{M}

Definition (Rule Redundancy)

A rule R is redundant in \mathcal{M} iff $\mathcal{M} \setminus R$ is equivalent to \mathcal{M}

Notations

- $\Delta(o)$: set of rules with outcome o
- Suppose that $\Delta(o) = \{R_1, \dots, R_z\} \cup \{R\}$
- Denote by $Rest = L_1 \vee \dots \vee L_z$

Rule Redundancy

Definition (Rule Redundancy)

A rule R is redundant in \mathcal{M} iff $\mathcal{M} \setminus R$ is equivalent to \mathcal{M}

Notations

- $\Delta(o)$: set of rules with outcome o
- Suppose that $\Delta(o) = \{R_1, \dots, R_z\} \cup \{R\}$
- Denote by $Rest = L_1 \vee \dots \vee L_z$

Proposition (Rule Redundancy Check)

A rule R is redundant in \mathcal{M} iff $\mathcal{B} \wedge L \models Rest$

Literal Redundancy

Notation

- Let $l \in L$. We denote by \mathcal{M}_l the model where l is removed from L

Notation

- Let $l \in L$. We denote by \mathcal{M}_l the model where l is removed from L

Definition (Literal Redundancy)

A literal l is redundant in L iff $l \in L$ and \mathcal{M}_l is equivalent to \mathcal{M} .

Example

- $R_1 = (salary > 0) \wedge (size \neq 140) \wedge (age > 10) \wedge (color = blue) \wedge (weight > 80) \implies 1$
- $(age > 10)$ is locally redundant in R_1 .

Example

- $R_1 = (salary > 0) \wedge (size \neq 140) \wedge (age > 10) \wedge (color = blue) \wedge (weight > 80) \implies 1$
- $(age > 10)$ is locally redundant in R_1 .

Lemma (Local Redundancy)

If $l \in L$ and $\mathcal{B} \wedge L \setminus \{l\} \models l$ then l is redundant in L .

Global Redundancy: Example

- $R_1 = (salary > 0) \wedge (size \neq 140) \wedge (age > 10) \wedge (color = blue) \wedge (weight > 80) \implies 1$
- $R_2 = (salary > 0) \wedge (size = 140) \implies 1$
- ...

$(size \neq 140)$ is globally redundant in R_1 :

- $Flip_{(size \neq 140)} = (salary > 0) \wedge (size = 140) \wedge (age > 10) \wedge (color = blue) \wedge (weight > 80)$
- $\mathcal{B} \wedge Flip_{(size \neq 140)} \models (salary > 0) \wedge (size = 140)$

Notation

- $\text{Flip}_I = L \cup \{\neg I\} \setminus \{I\}$

Notation

- $\text{Flip}_I = L \cup \{\neg I\} \setminus \{I\}$

Lemma (Global Redundancy)

If I is not locally redundant in L and $\mathcal{B} \wedge \text{Flip}_I \models \text{Rest}$, then I is redundant in L

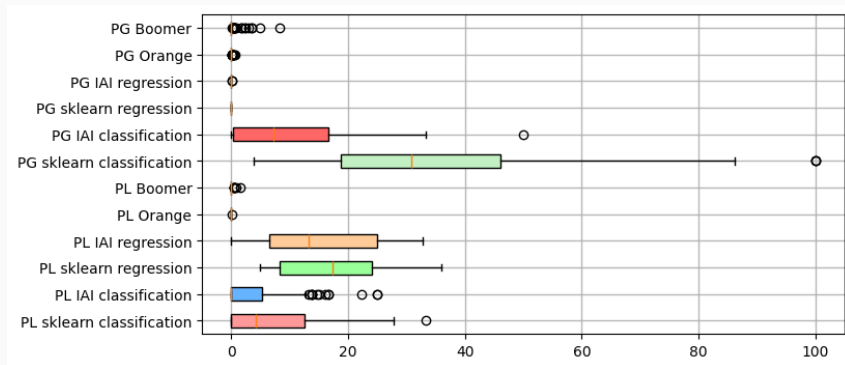
Experimental Results: Redundancy

Experimental Setting

- Scikit-learn and Interpretable AI for learning classification and regression decision trees
- Boomer to learn ensembles of boosted rules
- Orange v3 to learn decision sets for classification
- Diverse datasets from UCI ML repository with diverse characteristics
- Background constraints \mathcal{B} that enforce domain coherence between the features.
- SAT calls with PySAT
- One-hour time limit.
- All experiments ran on Apple M1 Pro (32 GB)

Frequency of Redundancies

Figure 2: Frequency of Literal Redundancy. PL (respectively PG) is the percentage of locally (respectively globally) redundant literals



Experimental Results: Overlaps

Overlapping Anchor Explanations

Learner	Dataset	Train	Test	# Explanations	# Overlap
xgboost	recidivism	92.39	74.33	333	87
randomforest	recidivism	93.52	75.46	321	65
logistic	recidivism	62.59	60.00	196	735
nn	recidivism	87.47	71.49	341	150
xgboost	lending	90.10	82.89	260	384
randomforest	lending	91.25	83.60	278	207
logistic	lending	82.56	83.51	50	54
nn	lending	88.00	82.54	159	66
xgboost	adult	90.35	84.26	565	3195
randomforest	adult	93.52	85.60	558	2534
logistic	adult	83.00	82.98	378	2788
nn	adult	92.47	83.62	597	3212

Conclusions

Contributions

1. Introduce a new approach to identifying and removing redundant information in rule based models under background constraints
2. Establish a dichotomy in the nature of redundancy
3. Propose novel algorithms for mining overlapping rules and explanations
4. Provide empirical insights:
 - Redundancy is widespread in commonly used tools
 - Overlapping rules occur frequently
 - Anchor explanations often lack trustworthiness

What I didn't have time to cover

- How to generate the background constraints
- The impact of the background constraints
- The computational overhead
- The quality of Overlap
- The correlation between redundancy and prediction quality
- The relationship with formal explainability (abductive explanations)

Future Research

- Develop stronger benchmarks for background (user-defined) constraints. If you know examples, please share with us!
- Incorporate distance metrics into regression tasks
- Why do Boomer and tree ensembles leverage overlap more effectively than Orange?