Explaining Cautious Random Forests via Counterfactuals

Haifei Zhang, Benjamin Quost, Marie-Hélène Masson

Université de Technologie de Compiègne, France
Université de Picardie Jules Verne, France
Laboratoire Heudiasyc, UMR CNRS 7253

14 September, SMPS 2022, Valladolid-Spain

Plan

- Introduction
- Methods
- Results
- Conclusion

Conversation between a loan applicant and a chatbot

Chatbot : Dear applicant, considering risk control, our system cannot decide whether to approve your application or not according to your profile.

Applicant : What ? That's a confusing result. Why can't your system make a decision on my application ? What can I do to get my application approved ? Or, how can I avoid being rejected ?

Chatbot : We are very sorry for that However, you have two options, either make an appointment with a manager for a consultation, or our system will generate a solution for you that will allow your application to be approved.

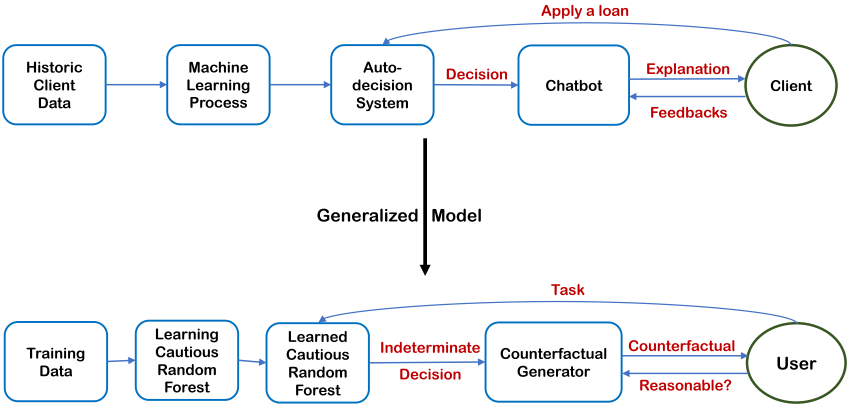
Conversation between a loan applicant and a chatbot

Applicant : For convenience, give me a feasible solution!

Chatbot : Okay, according to your profile, you currently have an income of €2,200 per month after taxes and have three credit cards. While keeping all other conditions unchanged, if you are able to reach an income of €2,400 and reduce your credit cards to two, your loan application will be approved.

Applicant : I see, this seems to be a reasonable solution.

Mechanism of the system



Cautious classifier

Disadvantage of traditional classifiers

Using traditional classifiers are risky when

- Scarce information : epistemic uncertainty.
- Conflicting information : aleatory uncertainty.

Purpose of cautious classifiers

- Modeling both aleatory and epistemic uncertainty in the reasoning process.
- Providing imprecise decisions, such as set-valued classification predictions and interval-valued regression values.

Cautious Random Forest is such kind of classifier!

Imprecise decision trees induced by IDM

Information about a decision region

1. A forest $F = \{C^1 \dots C^t \dots C^T\}$ has been trained for a binary classification problem, i.e., $y \in \{0, 1\}$.
2. A test instance x falls into a decision region that is the intersection of regions provided by trees, note as $R = \bigcap_{t=1\dots T} R^t$
3. In each region R^t, n^t and N^t count the number of class 1 and total training samples.

Imprecise Dirichlet Model (IDM) Intervals [5]

The interval-valued estimation $I_1^t = [\underline{p}_1^t, \overline{p}_1^t]$ of $p_1^t = Pr(Y = 1|x, C_t)$ is :

$$I_1^t = \left[\frac{n^t}{N^t + s}, \frac{n^t + s}{N^t + s}\right] \quad t = 1 \dots T$$

Aggregation of IDM intervals and decision making

Tree aggregation

For an instance x, each interval I_1^t provided by a tree is regarded as a focal element associated with a mass m_t on the interval $[0,1]$ [6]. The belief and plausibility of the event $Pr(Y=1|x) \geq 0.5$ can be defined as :

$$bel_1(x) = bel(Pr(Y=1|x) \in [0.5,1]) = \sum_{I_1^t \subseteq [0.5,1]} m_t = \sum_{t=1}^{T} m_t \mathbb{1}(\underline{p}_1^t \geq 0.5),$$

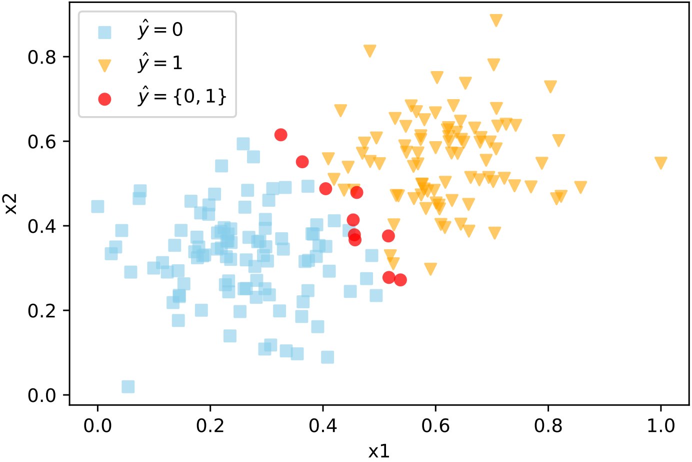
$$pl_1(x) = pl(Pr(Y=1|x) \in]0.5,1]) = \sum_{I_1^t \cap]0.5,1] \neq \emptyset} m_t = \sum_{t=1}^{T} m_t \mathbb{1}(\overline{p}_1^t > 0.5),$$

where $\mathbb{1}(\cdot)$ is the indicator function.

Decision strategy : interval dominance

$$\hat{y}_i = \begin{cases} 0 & \text{if } pl_{i,1} < 0.5 \\ 1 & \text{if } bel_{i,1} \geq 0.5 \\ \{0,1\} & \text{otherwise} \end{cases}$$

A typical example of cautious random forest on 2D data



Definition of counterfactual explanation

Counterfactual explanations are minimal alterations (exist in training data or synthesized) of an original query instance x leading to different predictions [4].

Given a classifier f, a query instance $x \in \mathcal{X}$, and a desired prediction label $y' \in \mathcal{Y}$, we aim at efficiently computing x' by solving

$$x' = \arg\min_{z \in \mathcal{X}} dist(x, z) \text{ s.t. } f(z) = y',\ f(x) \neq y',$$

where $dist(\cdot)$ is a suitable distance measure (e.g., Euclidean) between instances.

Definition of our problem

Given a trained cautious random forest f on binary classification data and an instance x with $f(x) = \{0, 1\}$ (indeterminate), we will search counterfactuals x' and x'' defined as

$$x' = \arg\min_{z \in \mathcal{X}} dist(x, z) \text{ s.t. } f(z) = 0$$

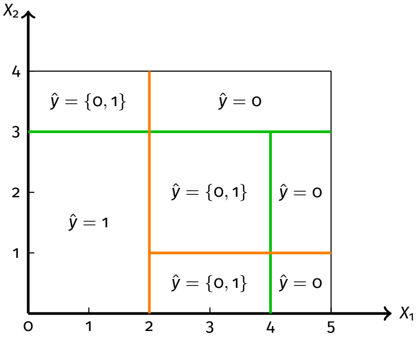
$$x'' = \arg\min_{z \in \mathcal{X}} dist(x, z) \text{ s.t. } f(z) = 1$$

where $dist(\cdot)$ is a suitable distance measure (e.g., Euclidean) between instances.

Counterfactual generation from random forests

Difficulties

1. Random forests are non-differential.
2. Trees in a forest are not totally independent.
3. The number of decision region (intersection of leaves) is exponential, i.e., for a forest of T trees, and each tree has L leaves, the complexity will be $O(L^T)$.



Regions of tree 1

	X_1	X_2
R_1 :	$\{[0, 2],$	$[0, 4]\}$
R_2 :	$\{]2, 5],$	$[0, 1]\}$
R_3 :	$\{]2, 5],$	$]1, 4]\}$

Regions of tree 2

	X_1	X_2
R_4 :	$\{[0, 4],$	$[0, 3]\}$
R_5 :	$\{]4, 5],$	$[0, 3]\}$
R_6 :	$\{[0, 5],$	$]3, 4]\}$

Counterfactual generation from cautious random forests

We adopted the method proposed by Blanchard [1] :

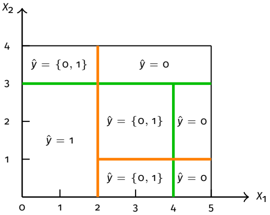
1. Convert random forest to a single search tree where each level represents the finest split of a feature of the input space.
2. Start from the region that contains x and explore nearby regions R with initial upper distance $d_{sup} = +\infty$,
   - if a region leads to candidate counterfactual (of desirable prediction and $d(x, R) < d_{sup}$), then generate counterfactual in the region and update d_{sup} ;
   - if not, go backwards in the search tree, dimension by dimension.

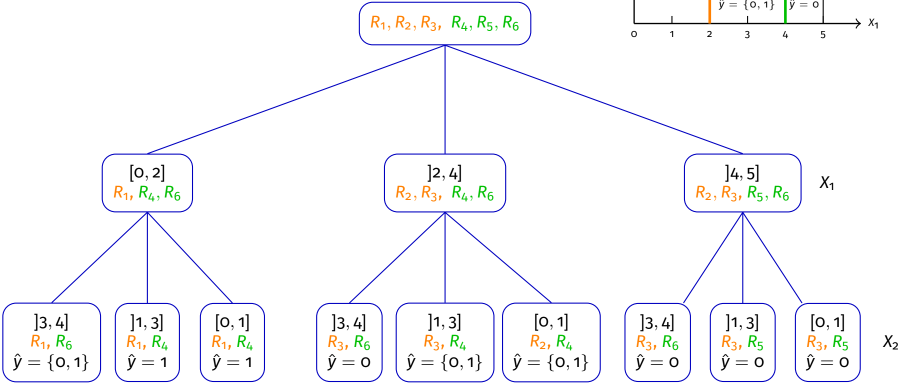
Reegions of tree 1

	X_1	X_2
R_1 :	$\{[0,2],$	$[0,4]\}$
R_2 :	$\{]2,5],$	$[0,1]\}$
R_3 :	$\{]2,5],$	$]1,4]\}$

Regions of tree 2

	X_1	X_2
R_4 :	$\{[0,4],$	$[0,3]\}$
R_5 :	$\{]4,5],$	$[0,3]\}$
R_6 :	$\{[0,5],$	$]3,4]\}$



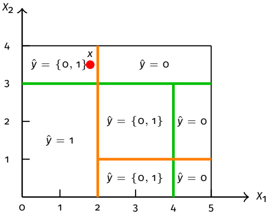


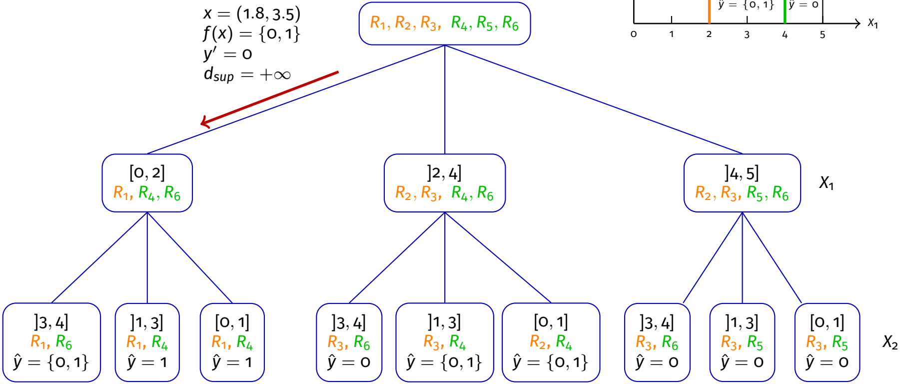
Reegions of tree 1

	X_1	X_2
R_1 :	$\{[0,2],$	$[0,4]\}$
R_2 :	$\{]2,5],$	$[0,1]\}$
R_3 :	$\{]2,5],$	$]1,4]\}$

Regions of tree 2

	X_1	X_2
R_4 :	$\{[0,4],$	$[0,3]\}$
R_5 :	$\{]4,5],$	$[0,3]\}$
R_6 :	$\{[0,5],$	$]3,4]\}$



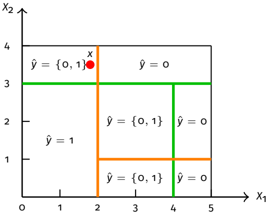


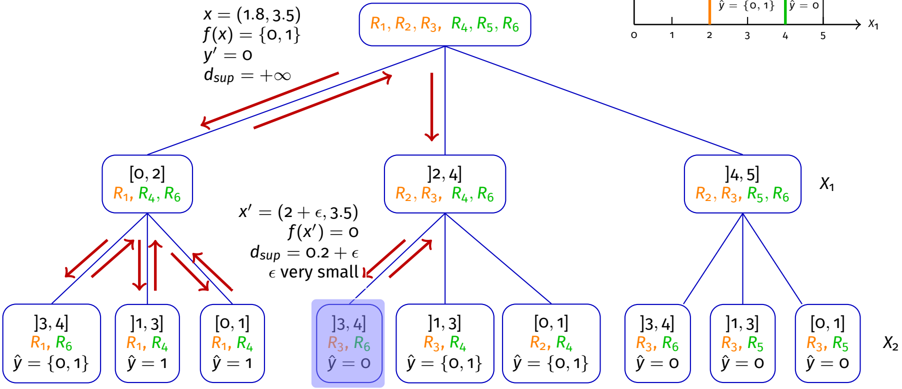
Reegions of tree 1

	X_1	X_2
R_1 :	{[0, 2],	[0, 4]}
R_2 :	{]2, 5],	[0, 1]}
R_3 :	{]2, 5],	]1, 4]}

Regions of tree 2

	X_1	X_2
R_4 :	{[0, 4],	[0, 3]}
R_5 :	{]4, 5],	[0, 3]}
R_6 :	{[0, 5],	]3, 4]}



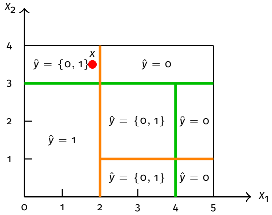


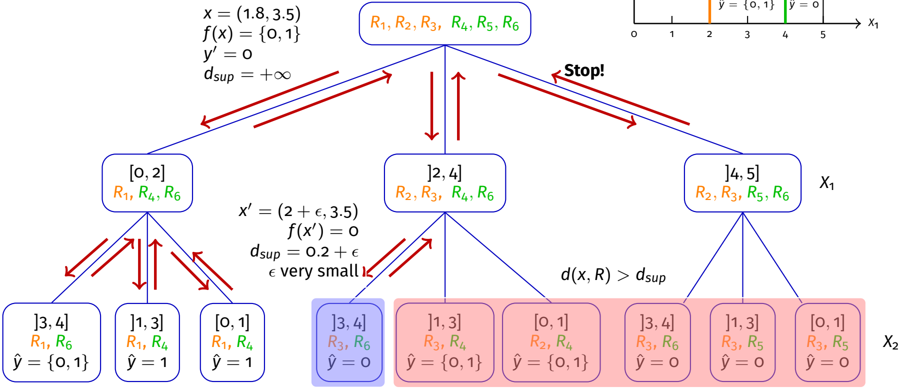
Reegions of tree 1

	X_1	X_2
R_1 :	$\{[0,2],$	$[0,4]\}$
R_2 :	$\{]2,5],$	$[0,1]\}$
R_3 :	$\{]2,5],$	$]1,4]\}$

Regions of tree 2

	X_1	X_2
R_4 :	$\{[0,4],$	$[0,3]\}$
R_5 :	$\{]4,5],$	$[0,3]\}$
R_6 :	$\{[0,5],$	$]3,4]\}$





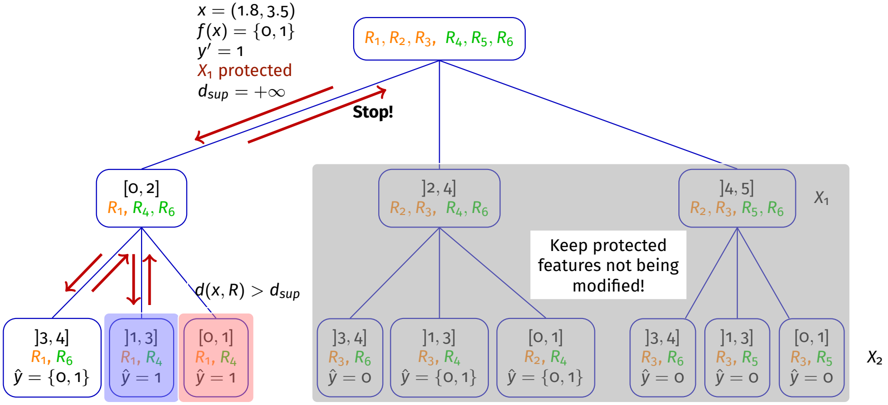
Efficiency issues

The branch-and-bound search method has two impact factors :

1. The depth of the search tree, i.e., how many features are mutable ?
2. The width of each level, i.e., how far should we explore for each feature ?

Proposed ssolution for efficiency issues 1

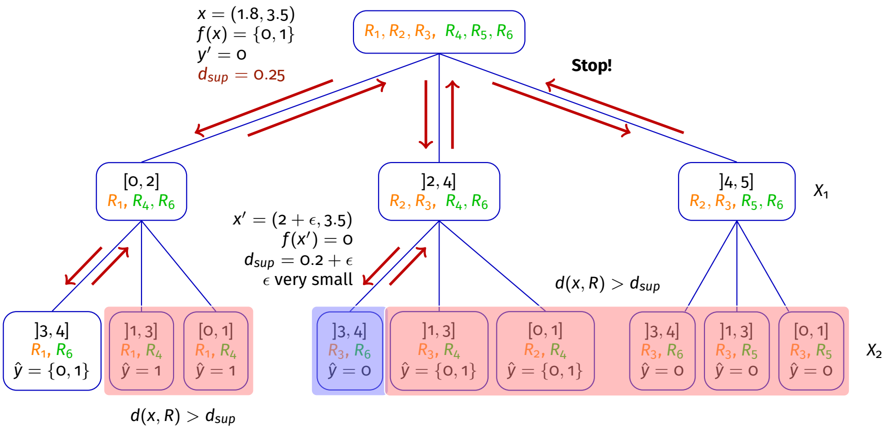
Reduced the depth of the search tree by introducing protected (immutable) features.



Here, if $y' = 0$, there will be no desirable and feasible counterfactual!

Proposed solution for efficiency issues 2

Control the width of each level of the search tree can be controlled by initializing a smaller upper distance d_{sup}, rather than $+\infty$.



One-dimensional Change CounterFactual (OCCF)

In explainable machine learning, Individual Conditional Expectation (ICE) plots display one line per instance that shows how the instance's prediction changes when only one feature changes [2].

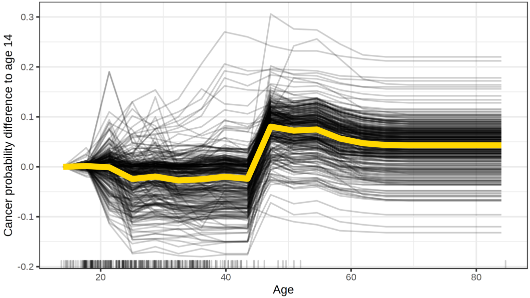


Figure – Example of ICE plot from Christoph Molnar's book "Interpretable ML" [3]

One-dimensional Change CounterFactual (OCCF)

Inspired by ICE, we proposed in our work to initial d_{sup} by the nearest counterfactual that changes only one feature,

$$d_{sup} = \min_{x' \in \mathcal{X}} dist(x, x'), \text{ s.t. } hd(x, x') = 1,$$

where *dist* is a selected distance measurement, e.g., Euclidean, and *hd* is Hamming distance.

Efficiency of search

Table – Initial upper bound distance by different methods

Dataset	MO	PF+MO	OCCF	PF+OCCF
Compas	0.078	0.134	**0.040**	0.058
Heloc	0.273	——	**0.011**	——
Pima	0.215	0.273	**0.034**	0.041
Wine	0.192	——	**0.060**	——

Table – Final counterfactual searching time cost

Dataset	MO	PF+MO	OCCF	PF+OCCF
Compas	1.091	0.421	0.580	**0.284**
Heloc	4.570	——	**1.274**	——
Pima	5.600	4.991	3.589	**3.277**
Wine	5.745	——	**4.667**	——

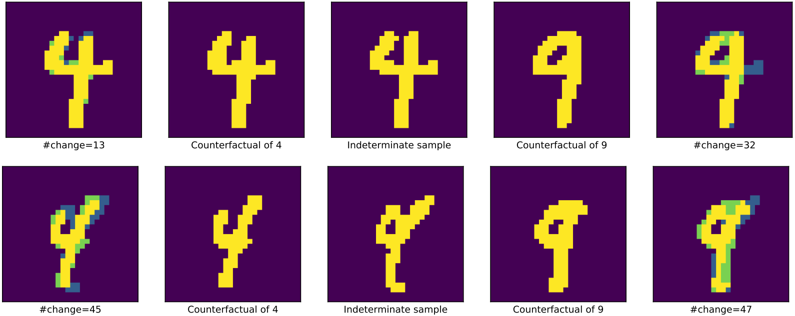
MO : Minimum Observable, searching for counterfactual in training set,
PF : Protected Features,
OCCF : One-dimensional Change CounterFactual.

Two-sided counterfactuals for an instance in Pima dataset

	PGs	Glucose	BP	ST	Insulin	BMI	DPF	Age
x	0	165	90	33	680	52.3	0.427	23
x_0	0	**154.5**↓	90	33	680	**47.7**↓	0.427	23
x_1	0	**165.5**↑	90	33	680	52.3	0.427	23

PGs : Pregnancy times, BP : Blood pressure, ST : Skin thickness,
BMI : Body mass index, DPF : Diabetes pedigree function.
PGs, DPF and Age are protected features.

Four or Nine ?



Left- and right-most images display pixels to be added (green) and to be deleted (blue) in order to obtain the counterfactual.

Conclusion

Summary of this work

- We proposed to use counterfactuals to explain the imprecision of cautious random forests, which is intuitive and easy to understand.
- We proposed a new counterfactual initialization method (OCCF) to speed up the process of generating counterfactuals from random forests.

Conclusion

Summary of this work

- We proposed to use counterfactuals to explain the imprecision of cautious random forests, which is intuitive and easy to understand.
- We proposed a new counterfactual initialization method (OCCF) to speed up the process of generating counterfactuals from random forests.

Future directions

- Study how to quickly generate counterfactuals for samples far from the classification boundary.
- Consider how to solve the counterfactual plausibility problem.

Conclusion

Summary of this work

- We proposed to use counterfactuals to explain the imprecision of cautious random forests, which is intuitive and easy to understand.
- We proposed a new counterfactual initialization method (OCCF) to speed up the process of generating counterfactuals from random forests.

Future directions

- Study how to quickly generate counterfactuals for samples far from the classification boundary.
- Consider how to solve the counterfactual plausibility problem.

Thanks for your attention!

References

[1] Pierre Blanchart. « An exact counterfactual-example-based approach to tree-ensemble models interpretability ». In : *arXiv preprint arXiv :2105.14820* (2021).

[2] Alex Goldstein, Adam Kapelner, Justin Bleich et Emil Pitkin. « Peeking inside the black box : Visualizing statistical learning with plots of individual conditional expectation ». In : *journal of Computational and Graphical Statistics* 24.1 (2015), p. 44-65.

[3] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[4] Sandra Wachter, Brent Mittelstadt et Chris Russell. « Counterfactual explanations without opening the black box : Automated decisions and the GDPR ». In : *Harv. JL & Tech.* 31 (2017), p. 841.

References

[5] Peter WALLEY. « Inferences from multinomial data : learning about a bag of marbles ». In : *Journal of the Royal Statistical Society : Series B (Methodological)* 58.1 (1996), p. 3-34.

[6] Haifei ZHANG, Benjamin QUOST et Marie-Helène MASSON. « Cautious Random Forests : a new decision strategy and some experiments ». In : *International Symposium on Imprecise Probability : Theories and Applications*. PMLR. 2021, p. 369-372.