Explaining Cautious Random Forests via Counterfactuals

Haifei Zhang, Benjamin Quost, Marie-Hélène Masson

Abstract Cautious random forests are designed to make indeterminate decisions when tree outputs are conflicting. Since indeterminacy has a cost, it seems desirable to highlight why a precise decision could not be made for an instance, or which minimal modifications can be made to the instance so that the decision becomes a single class. In this paper, we apply an efficient extractor to generate determinate counterfactual examples of different classes, which are used to explain indeterminacy. We evaluate the efficiency of our strategy on different datasets and we illustrate it on two simple case studies involving both tabular and image data.

1 Introduction

Machine learning models now achieve high performances in many fields such as medical diagnosis, recommendation systems, image and speech recognition. The outputs of these models are traditionally precise: in a classification problem, they consist in a single class for a given instance. However, when training data are scarce, or when mistakes have a very high cost, cautious classifiers can alternatively be used to provide set-valued decisions rather than single classes and thus control the risk. Cautious random forests (CRF) (Zhang et al., 2021) are one of those classifiers. A CRF combines the classical random forest (RF) strategy (Breiman, 2001), the Imprecise Dirichlet Model (IDM) (Walley, 1996) and the theory of belief functions (Shafer, 1976). The

Haifei Zhang (✉), e-mail: haifei.zhang@hds.utc.fr
Benjamin Quost (✉), e-mail: benjamin.quost@hds.utc.fr
UMR CNRS 7253 Heudiasyc, Université de Technologie de Compiègne, 60200, France

Marie-Hélène Masson (✉), e-mail: mylene.masson@hds.utc.fr
UMR CNRS 7253 Heudiasyc, Université de Picardie Jules Verne, IUT de l'Oise, Beauvais, 60000, France

major difference with a classical RF is that an indeterminate decision can be reached in presence of both epistemic uncertainty (when the tree outputs are based on scarce information) and aleatoric uncertainty (the conflict between these outputs is high), which typically happens near decision boundaries. Making imprecise predictions has a cost, since indeterminacy must be resolved via further analysis. Therefore, it seems crucial to understand what led to an undetermined decision, and what could be done to change it into a determinate one. Such questions fall under the emerging topic of explainable machine learning (Molnar, 2019). In this paper, we address the second problem using counterfactual explanations (Wachter et al., 2017), which provide clear and intuitive explanations for turning an original instance x into a modified one x' in a minimal way, so that $f(x')$ corresponds to a desired prediction $y' \neq f(x)$. Our approach is inspired by the one proposed by Blanchart (2021), specifically developed for tree ensembles. Our contributions consist in improving the efficiency of the procedure, and in exploiting counterfactuals for explaining indeterminate CRF outputs. Their benefits for explaining indeterminacy are illustrated by experimental results, in particular via two case studies.

The paper is structured as follows. In Section 2, we recall general background knowledge on cautious random forests and counterfactual explanations. Their application to explaining indeterminacy are discussed in Section 3. Section 4 details the experiments and discusses the results. A short conclusion is drawn in Section 5.

2 Background

2.1 Cautious Random Forests

Cautious random forests (CRF) have been proposed as an alternative to precise random forests, so as to make decisions from scarce data. In a binary classification problem, for each test instance x, each tree t in the forest provide pieces of evidence about its actual class $Y \in \{1, 0\}$ in the form of lower and upper bounds $\underline{p}_1^t(x)$ and $\overline{p}_1^t(x)$ over the posterior probability $\Pr(Y = 1|x)$. These bounds are obtained using the Imprecise Dirichlet Model, and reflect the estimation uncertainty due to the lack of training data. These intervals can be pooled using the theory of belief functions, by computing the belief and plausibility $bel(Y = 1|x)$ and $pl(Y = 1|x)$, which can then be used in a cautious decision-making process such as interval dominance, possibly resulting in indeterminate decisions (Zhang et al., 2021).

As can be seen in Fig. 1, imprecision occurs principally around the decision boundaries, where tree leaves are prone to contain few instances and tree outputs are often conflicting with each other.

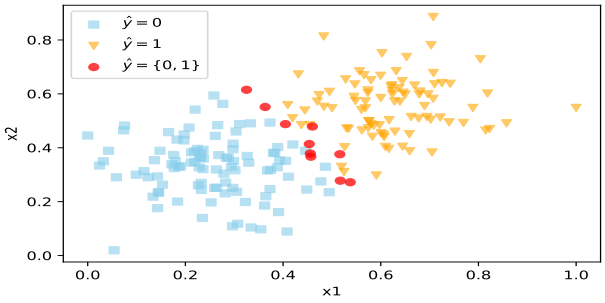

Fig. 1 Predictions of a CRF with $T = 100$ trees.

2.2 Counterfactual-based explanations

Explainable artificial intelligence is an emerging field of artificial intelligence that helps humans to understand the outputs of machine learning algorithms. Counterfactuals (CF) (Wachter et al., 2017) are local and example-based explanations, which can be seen as minimal alterations of an original query instance x leading to different decisions. Such examples can either be queried for in the training set, or synthesized. 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

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

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

Many methods have been designed to solve (1) exactly or approximately, such as selecting the most similar sample in the training set, or creating a virtual sample by optimizing a loss function (for differentiable models), searching CFs by a heuristic strategy, or approximating the model at hand (e.g. by a decision tree) so as to simplify the search of a CF (Guidotti, 2022). Besides the inherent complexity of counterfactual generation algorithms, additional challenges make designing an actionable decision process difficult (Verma et al., 2020), such as protecting some attributes or immutable features (such as gender, ethnicity, etc), restricting the number of modified features (sparsity), and generating plausible (or realistic) CFs.

3 Explaining imprecision using counterfactuals

This work proposes to apply CF explanations to cautious binary classification: given an instance x with indeterminate prediction $f(x) = \{0, 1\}$, we want to identify its two minimal modifications x^1 and x^0 such that $f(x^1) = \{1\}$ and $f(x^0) = \{0\}$. These two synthetic examples will not only reveal the features

which should be modified to remove indeterminacy, but also to which extent they should be modified so as to reach a precise decision.

3.1 Extracting determinate counterfactuals

We propose to extract counterfactuals using the geometrical method of Blanchart (2021), which explicitly computes the smallest decision regions of a tree ensemble model, and generates the closest virtual CF in terms of Euclidean distance. A random forest separates the input space $\mathcal{X}$ into decision regions, each of which is itself the intersection of T decision regions provided by the T trees in the forest. Computing the optimal and exact CF with desired class y' for an instance x requires to explore all decision regions of the forest, the complexity of which is exponential. This exhaustive search is thus intractable for high-dimensional data or forests with deep trees. Blanchart (2021) proposed a branch-and-bound strategy to search for CFs only around x, by ignoring decision regions R such that $d(x, R) > d_{\max}$, with $d_{\max}$ (initialized to positive infinity) the distance to the current counterfactual found during the extraction procedure. We refer the reader to this reference for further information.

3.2 Region filtering and counterfactual initialization

Given the complexity of determining a CF x' for a given query instance x with indeterminate decision $f(x) = \{1, 0\}$, we propose two amenities to speed up the procedure. These preliminary steps make it possible to drastically reduce the complexity of the search, as will be shown in Section 4.

1. Following a suggestion of Blanchart (2021), in presence of protected features, we filter out the regions that do not correspond to the same protected values as in x.
2. We use an alternative approach to "initialize" the CF search, i.e. to compute the first CF based upon which the initial distance threshold $d_{\max}$ will be determined, rather than positive infinity.

Embedding the filtering step mentioned above in any branch-and-bound procedure is straightforward. The initialization step is critical, since it determines the threshold $d_{\max}$ and therefore the number of regions to be explored.

The Minimum Observable (MO) approach, which selects the nearest instance x' with desired class y' in the training set, is commonly used for this purpose. However, in scarce regions of the input space, the distance between the query point and the closest training CF may be large. Even worse, when several protected features (PF) are considered, the approach may not give an initial CF which meets the requirements. Therefore, we propose a new

strategy to find an initial virtual CF, which we call One-dimensional Change CounterFactual (OCCF). In a nutshell, for a given x, OCCF aims at solving Eq. (1) with the additional constraint that x and z differ by one feature only. This problem can be quickly solved using individual conditional expectation (ICE) plots (Goldstein et al., 2015), which estimate how the probability (or the decision) $f(x)$ of a classifier varies according to a modification in x when all other values are fixed.

Note that in a random forest, for a query instance x, we need only consider a finite number of modifications of the value of a mutable feature $\mathcal{X}_d$, defined by the split values for this feature obtained across all trees. However, an OCCF may still not exist when several features are protected, although the experiments suggest that this is much less likely than with the MO approach. In this case, some constraints should be relaxed by "unprotecting" some immutable features.

4 Experimental results

4.1 Counterfactual extraction efficiency

In this experiment, we evaluate the efficiency of the proposed CF extraction procedure on four datasets. The number of trees in the ensemble is 50 for all datasets, and the maximal depth of the trees are respectively 10, 8, 7, and 14. The efficiency is evaluated in three ways: the number of regions to explore after filtering by different initialization approaches, the distance between the query point and the initial CF, and the elapsed time to extract all CFs. Note that Compas and Pima have one and four protected features, respectively, whereas no protected features were considered for Heloc and Wine.

Table 1 Average number of leaves to explore

Dataset	Original	PF	MO	PF+MO	OCCF	PF+OCCF
Compas	7236	2226.86	849.87	732.54	418.36	**305.36**
Heloc	8784	———	5268.12	———	**106.52**	———
Pima	2522	1081.27	1007.43	719.53	133.90	**128.77**
Wine	8949	———	3277.05	———	**761.38**	———

Tables 1 and 2 indicate that exploiting the protected features can help to reduce the amount of regions to explore, since it restricts searching the CFs to a feature subspace. Our proposed OCCF initialization can generate initial CFs which are much closer to the query point x compared to MO: as a consequence, we may filter out many more regions, and thus reduce the amount of time needed to reach the solution.

Table 2 Average distance from the query example to the initial counterfactual (left), and average elapsed time for searching the final counterfactual (right)

	Initial CF Distance				CF Searching Time (s)			
Dataset	MO	PF+MO	OCCF	PF+OCCF	MO	PF+MO	OCCF	PF+OCCF
Compas	0.078	0.134	**0.040**	0.058	1.091	0.421	0.580	**0.284**
Heloc	0.273	——	**0.011**	——	4.570	——	**1.274**	——
Pima	0.215	0.273	**0.034**	0.041	5.600	4.991	3.589	**3.277**
Wine	0.192	——	**0.060**	——	5.745	——	**4.667**	——

4.2 Case studies

Case 1: Pima

The Pima dataset can be used to predict whether a patient has diabetes or not, based on various measurements: Pregnancies (PGs): number of times pregnant; Glucose; Blood Pressure (BP); Skin Thickness (ST); Insulin: 2-Hour serum insulin (mu U/ml); BMI: body mass index; Diabetes Pedigree Function (DPF); Age. The class is $y = 0$ for a non-diabetic, $y = 1$ for a diabetic. Here, Age, number of pregnancies, DPF values, and Skin Thickness are difficult to change (considered as protected features), while Glucose, Insulin, BMI, and blood pressure are actionable (mutable) features. We chose Pima as an example because, as a medical dataset, explainability may have a great practical interest.

Table 3 Examples of counterfactual explanations from Pima dataset.

	PGs	Glucose	BP	ST	Insulin	BMI	DPF	Age
x_1	0	165	90	33	680	52.3	0.427	23
x_1^0	0	**154.5**$\downarrow$	90	33	680	**47.7**$\downarrow$	0.427	23
x_1^1	0	**165.5**$\uparrow$	90	33	680	52.3	0.427	23
x_2	1	122	90	51	220	49.7	0.325	31
x_2^0	1	**121.5**$\downarrow$	90	51	**128**$\downarrow$	**49.05**$\downarrow$	0.325	31
x_2^1	1	**126.5**$\uparrow$	90	51	220	49.7	0.325	31

In Table 3, two examples are provided for illustration. The query instance x_1 corresponds to a non-diabetic patient. First, note that x_1 is close to being classified as diabetic since the CF x_1^1 of this class is very close. This demonstrates that the cautious random forest can help managing the uncertainty arising from scarce data, by detecting instances for which the decision is uncertain and providing insights about their actual labels. Second, the non-diabetic CF x_1^0 suggests a possible way to maintain a healthy condition, i.e reducing BMI and the Glucose level. The query instance x_2 corresponds to

a diabetic patient. The indeterminacy comes from the Glucose feature, since we can get a correct prediction (diabetic) by only modifying its value. On the other hand, to obtain the non-diabetic CF x_2^0, an important decrease of Insulin is needed, which is coherent with the fact that high 2-Hour serum insulin levels are common for type-II diabetic patients.

Case 2: MNIST

MNIST is a large database of handwritten numbers containing about 60,000 training cases and 10,000 test cases. In our experiment, numbers of 4 and of 9 were selected and 40 principal components of the original data were extracted to train a CRF consisting of 50 trees of depth 10. We generated CFs, which we required to belong to the objective class with a belief of at least 0.75, so as to ensure that the instance is credible after applying the inverse PCA transformation. Generating CFs of a query instance helps understanding which parts of the image are responsible for the indeterminacy of the decision. This point is illustrated using two instances drawn in Figure 2. We can see how the two indeterminate examples (center) should be modified to be determinately classified as a "4" or as "9", and that these modifications make sense.

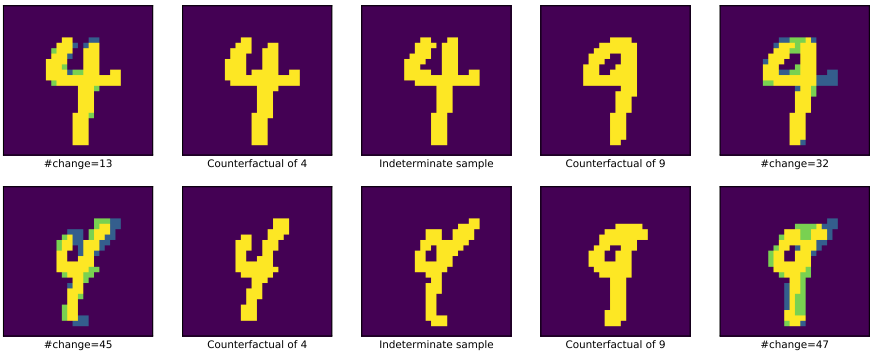

Fig. 2 Examples of indeterminate numbers (center) and corresponding counterfactuals of class 4 (left) and 9 (right). Left- and right-most images display pixels to be added (green) and to be deleted (blue) in order to obtain the counterfactual.

5 Conclusion

In this paper, we have proposed a procedure to extract CFs of indeterminate instances, i.e. for which no precise decision could be made, so as to interpret and explain the indeterminacy of the classifier. The algorithm presented in

this paper is specific to cautious random forests. It is based on an algorithm proposed in the case of precise CF extraction. Our modifications make it possible to filter the regions of the input space to be explored and to generate CFs closer to the query instance, thus speeding up the extraction process. This increased efficiency, as well as the usefulness of our approach for explaining the indecision of the classifier, has been demonstrated on several experiments. In future works, we will investigate how CFs can be used to estimate the importance of features and to identify regions of significant uncertainty in the feature space. We also plan to use CFs in an active learning process to reduce the indeterminacy of cautious classifiers.

References

Blanchart P (2021) An exact counterfactual-example-based approach to tree-ensemble models interpretability. arXiv preprint arXiv:210514820

Breiman L (2001) Random forests. Machine learning 45(1):5–32

Goldstein A, Kapelner A, Bleich J, Pitkin E (2015) Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics 24(1):44–65

Guidotti R (2022) Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery pp 1–55

Molnar C (2019) Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/

Shafer G (1976) A mathematical theory of evidence. Princeton university press

Verma S, Dickerson J, Hines K (2020) Counterfactual Explanations for Machine Learning: A Review. In: NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-Analyses (ML-RSA)

Wachter S, Mittelstadt B, Russell C (2017) Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harvard Journal of Law & Technology 31:841

Walley P (1996) Inferences from Multinomial Data: Learning About a Bag of Marbles. Journal of the Royal Statistical Society: Series B (Methodological) 58:3–34

Zhang H, Quost B, Masson MH (2021) Cautious random forests: a new decision strategy and some experiments. In: International Symposium on Imprecise Probability: Theories and Applications, PMLR, pp 369–372