BoXoR-C: Coherence and Directionality of Counterfactual Explanations

Antonio Maratea, Domenico Lugubre

Department of Science and Technologies University of Naples "Parthenope", 80143 Naples, Italy
antonio.maratea@uniparthenope.it

Aim of the work

Aim of the work

Given a set of counterfactual explanations for a given model, to identify through clustering regions where these explanations are stable and trustable, as well as regions where they are not.

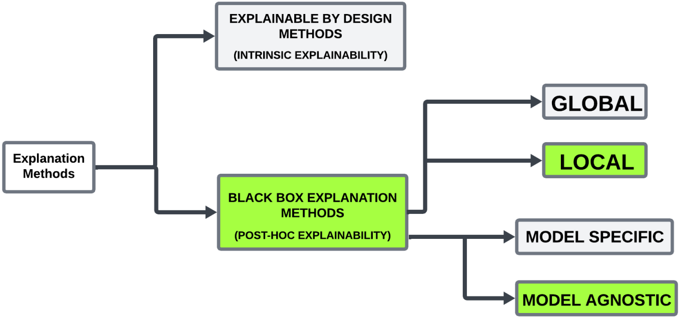
Outline

- XAI and counterfactuals
- Granularity, the Rashomon effect and other issues
- BoC-SoR vs BoXoR-C
- Tests
- Conclusions

Counterfactual definition

what-if

For an instance s_i, given a classifier and its predicted class l_i, its counterfactual is another instance c_j with a different label $l_j, l_i \neq l_j$ that has a minimal difference with respect to s_i.

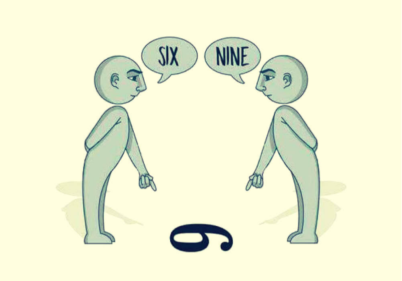


Issues for counterfactuals

- Granularity: counterfactuals are valid for a single instance;
- Variability: presence of several different counterfactuals for the same instance;
- Actionability, that is the actual feasibility of the suggested changes on the target variables;
- Directionality, that is the difference involved in reversing the change from one class to the other.

The Rashomon effect

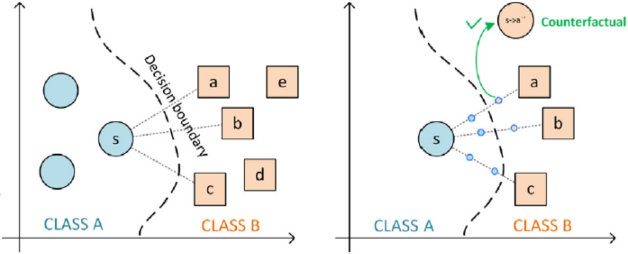
The presence of several different counterfactuals for the same instance, often contradicting each other, is called Rashomon effect.



It strongly limits the human trust in the counterfactuals due to the conflicting explanations, ambiguous causality and cherry picking.

BoC-SoR: Boundary Crossing Solo Ratio

The Boundary Crossing Solo Ratio is the baseline method on which we build up. It performs a feature relevance analysis on a per-istance scale.



Alfeo A. L. *et al.*, "From local counterfactuals to global feature importance: efficient, robust, and model-agnostic explanations for brain connectivity networks", Computer Methods and Programs in Biomedicine, Volume 236, 2023, 107550

BoXoR-C: clustering counterfactuals

Given a binary problem, the original class O and the opposite class C, for each instance $o \in O$ its closest instance of the other class is $c_{nn_o} \in C$.

Steps

- The set B of boundary instances is defined as follows:

$$B = \{b \in O \mid \text{dist}(b, c_{nn_b}) < \text{perc}(th, D)\},\ D = \{\text{dist}(o, c_{nn_o}) \mid o \in O\};$$

- The B instances are grouped into k clusters, $C_1, \ldots, C_k$;
- $\forall s_i \in C_j$ the counterfactual c_i is generated with BoC-SoR;
- $\forall c_i$ the subset of **relevant features** $\mathcal{F}^i_{s_i \rightarrow c_i}$ is extracted;
- The **binary matrix** $E_j \in \{0, 1\}^{n_j \times f}$ with $E_{mn} = 1$ if feature n is relevant for instance m is built and clustered into w clusters $\{G_1, \ldots, G_w\}$.

BoXoR-C: the proposed Method

1. Cluster the boundary points B into k clusters $\{C_1, \dots, C_k\}$

$$d_{\text{Euclidean}}(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

BoC-SoR

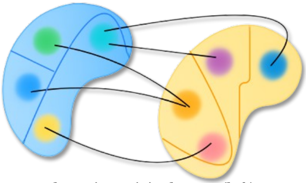
2. For each $s_i \in C_j$, generate the counterfactual c_i
   findCF(s_i)

3. Extract the subset of relevant features $\mathcal{F}^i_{s_i \to c_i}$
   relevantFeatures(s_i, c_i)

4. Construct the binary matrix $E_j \in \{0, 1\}^{n_j \times f}$ with $E_{mn} = 1$ if feature n is relevant for instance m

5. Cluster on $\mathbf{E}$ into w clusters $\{G_1, \dots, G_w\}$

$$d_{\text{Jaccard}}(E_i, E_j) = 1 - \frac{|E_i \cap E_j|}{|E_i \cup E_j|}$$



clustering original space (left)
and in the explanation space (right)

Validation

Within cluster variability indices

- The Shannon entropy is computed for each cluster as a global measure of diversity in explanations;
- For each feature F_n its prevalence $p_n = \sum_m^{n_j} \frac{E_{mn}}{n_j}$ is computed;
- The binary variance $\mathrm{Var_n} = p_n(1 - p_n)$ of each feature F_n within the cluster is also computed.

Clustering agreement measures

- Adjusted Rand Index (ARI)
- Normalized Mutual Information (NMI),
- Homogeneity, Completeness, and V-measure.

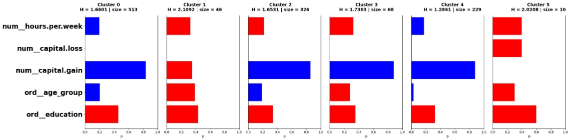
Adult Income, from class 0 to 1

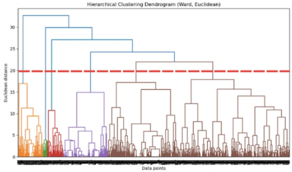
ADULT INCOME DATASET

Age	Educ.	Occup.	Workcl.	Sex	Hrs	Mar.	Relat.	Race	Income
39	Bachelors	Exec-mgr.	Private	Male	40	Never-mar.	Not-in-fam.	White	$\leq$ 50k
50	HS-grad	Mach-insp.	Self-emp	Male	45	Married	Husband	Black	$>$ 50k
...	...	...	...	...	...	...	...	...	...

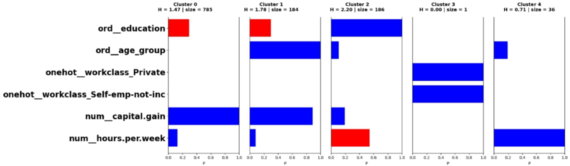
Source: UCI Adult Dataset https://archive.ics.uci.edu/ml/datasets/adult

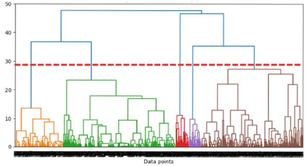
Original feature space





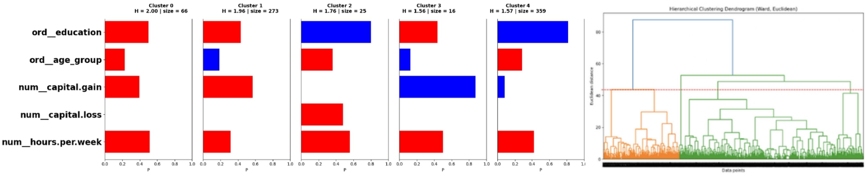
Explanation space



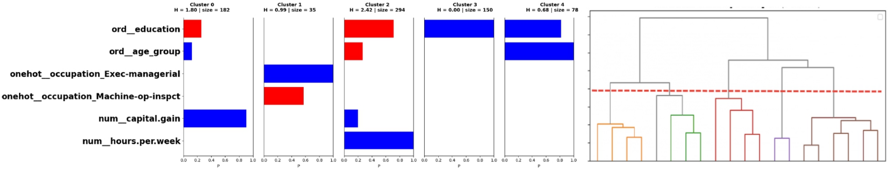


Adult Income, from class 1 to 0

Original feature space



Explanation space

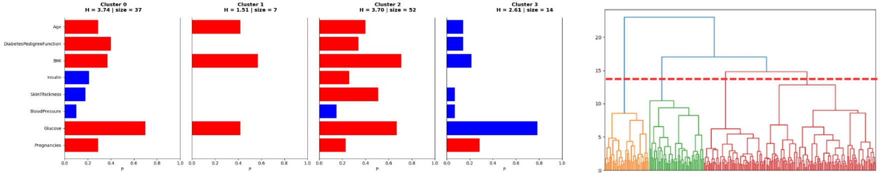


Diabete, from class 0 to 1

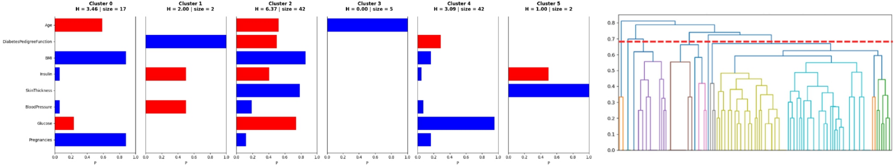
Diabetes dataset

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	138	70	35	120	33.6	0.127	45	1
5	155	72	28	180	31.0	0.489	54	0
...	...	...	...	...	...	...	...	...

Original feature space

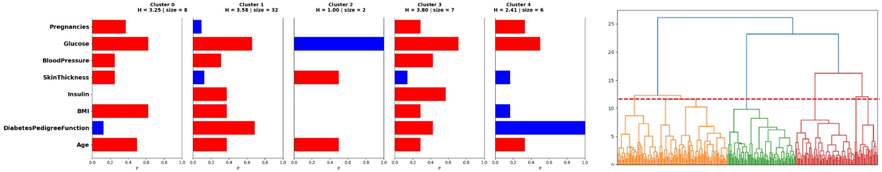


Explanation space

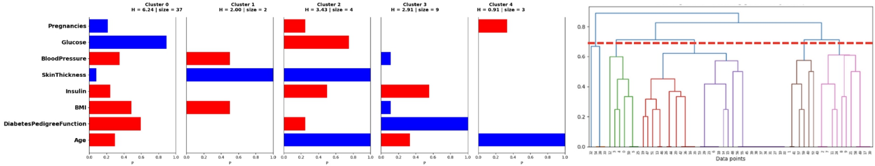


Diabete, from class 1 to 0

Original feature space



Explanation space



Conclusions

- A method to characterize ”safe” regions, where the counterfactuals explanations and the feature involved are stable and with a limited Rashomon effect and to recognize ”unsafe” regions, where the counterfactuals explanations and the feature involved change randomly, has been proposed;
- Clustering has been used in both the original feature space and the explanation space, accounting for label directionality;
- Results on real data are encouraging and confirm the viability and effectiveness of the method.

Thank you!