Security of Counterfactual Explanations

based on

"The Privacy Issue of Counterfactual Explanations: Explanation Linkage Attacks"

by S. Goethals, K. Sorensen and D. Martens

Mikołaj Spytek

MI2.AI Seminar, Warsaw, 15th April 2024

MI AI

Wydział Matematyki
i Nauk Informacyjnych
Politechnika Warszawska

What do the authors promise? (a.k.a. agenda)

- a description of a new **explanation linkage attack** that can be applied when **counterfactual explanations are based on real instances** from the training set,
- a solution in the form of **k-anonymous counterfactual explanations,**
- an evaluation of how anonymizing explanations **decreases their quality,**
- a discussion of the **trade-off** between **transparency, fairness** and **privacy.**

What kinds of counterfactuals are considered?

"counterfactual explanations [...] are defined as the smallest change to feature values of an instance that alters its prediction."

- **Explanation linkage attacks** are possible, when the counterfactual algorithm uses instance-based strategies to find the counterfactual explanations a.k.a. **nearest unlike neighbor.** (*note:* **plausibility**)

- Methods which produce vulnerable counterfactuals: NICE, WIT with NNCE, FACE, and certain settings of

How can explanatory variables be divided?

- **identifiers** - e.g., name, phone need to be suppressed always, often not fed to the models as they don't have predictive power
- **quasi identifiers** - cannot directly identify a person, but their combination might (e.g., zip code + year of birth)
- **private attributes** - attributes that are not publicly known

Identifier	**Quasi-identifiers**			**Private-attributes**		**Model pred.**
Name	Age	Gender	City	Salary	Relationship	Credit decision
Lisa	21	F	Brussels	$50k	Single	Reject

Table: An example of a factual instance: Lisa.

What is the goal of an adversary?

An adversary tries to **get access to** a user's **private attributes**, for example by asking for counterfactual explanations.

We also assume that all quasi-identifiers are public knowledge

How can explanatory variables be divided?

Name	Age	Gender	City	Salary	Relationship	Credit decision
Lisa	21	F	Brussels	$50k	Single	Reject
Alfred	25	M	Antwerp	$40k	Separated	Reject
Derek	47	M	Brussels	$100k	Married	Accept
Fiona	24	F	Antwerp	$60k	Single	Accept
Gina	27	F	Antwerp	$80k	Married	Accept

Table: Training set

If you were **3 years older**, lived in **Antwerp** and your income was **$10k higher**, then you would have **received** the loan

What do we know now?

Lisa (or an adversary pretending to be her) now knows that **Fiona** earns **$60k** and is **single**.

Based on her Lisa's own attributes and the counterfactual.

*** We assumed that knowing all quasi-identifiers can directly identify Fiona (e.g., by looking on social media, or voter's registration).**

k-anonymity

A counterfactual instance is considered to be **k-anonymous** if the combination of quasi-identifiers can belong to **at least *k* individuals** in the training set

2-anonymous example

Name	Age	Gender	City	Salary	Relationship	Credit decision
Lisa	21	F	Brussels	$50k	Single	Reject
Alfred	25	M	Antwerp	$40k	Separated	Reject
Derek	47	M	Brussels	$100k	Married	Accept
Fiona	24	F	Antwerp	$60k	Single	Accept
Gina	27	F	Antwerp	$80k	Married	Accept

Table: Training set

If you were **3-6 years older**, lived in **Antwerp** and your income was **$10k higher**, then you would have **received** the loan

An example of a 2-anonymous explanation. The counterfactual does not distinguish between Fiona and Gina in terms of quasi-identifiers

Evaluation metrics

- **Degree of privacy:**
  k - number of observations, between which it is impossible to distinguish based on quasi-identifiers.

- **Counterfactual validity:**
  pureness - in what percentage of samples from our counterfactual range, the decision is actually counterfactual

$$\text{pureness} = \frac{\#\text{ of value combinations with desired prediction outcome}}{\#\text{ of value combinations}}$$

Pureness example

Age	Gender	City	Salary	Relationship	Credit decision
24	F	Antwerp	$60k	Single	Accept
25	F	Antwerp	$60k	Single	Accept
26	F	Antwerp	$60k	Single	Reject
27	F	Antwerp	$60k	Single	Reject

Table: All possible values of attributes occurring in the data from the 2-anonymous explanation.

pureness = 2/4

*In practice, the authors do not estimate the exact pureness, by querying for all possible combinations, rather they **sample only 100 points.**

Evaluation metrics cont'd

- **Loss in information value:**
  Normalized Certainty Penalty

$$NCP_{A_{num}}(G) = \frac{\max^G_{A_{num}} - \min^G_{A_{num}}}{\max_{A_{num}} - \min_{A_{num}}}$$

$$\begin{cases} NCP_{A_{cat}}(G) = 0, & \text{if} |A^G| = 1 \\ NCP_{A_{cat}}(G) = \dfrac{|A^G|}{|A|}, & \text{otherwise} \end{cases}$$

$$NCP(G) = \sum_{i=1}^{d} w_i NCP_{A_i}(G)$$

Methodology

- split the data 60-40
- fit random forest and NICE counterfactual algorithm **on train**
- predict and obtain counterfactuals **on test**
- use the CF-K algorithm to anonymize the explanations

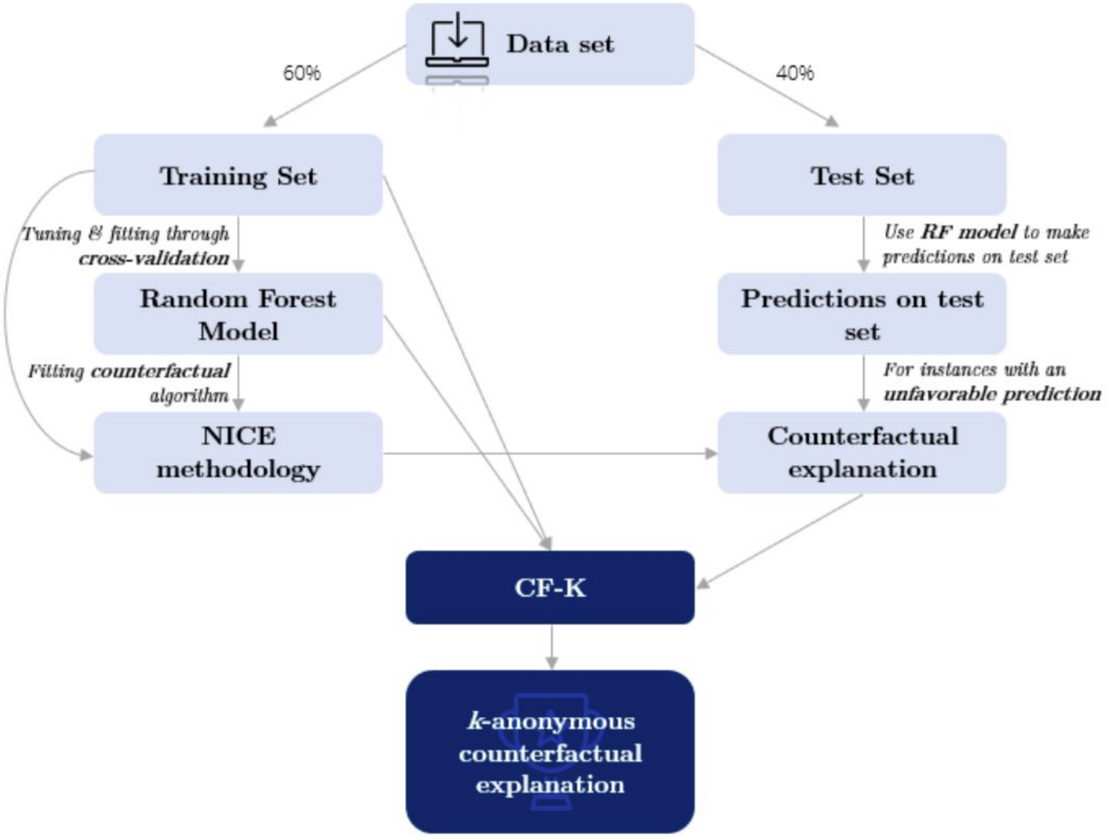

Figure: Methodology diagram. Source: original paper.

The CF-K Algorithm

- **Phase 1:** (construct a greedy randomized solution)

  Check if the current solution is at least **k-anonymous**. If it is, move to the next step. If not, **generate** a list of α closest counterfactual **neighbors**, and randomly select one of them. Then **generalize** the current solution with this new observation.

Loop until the found solution is k-anonymous

The CF-K Algorithm

- **Phase 2:** (perform a local search)

  Try to change the current solution by **slightly altering** the **quasi-identifier** values.

Terminate when the solution is no longer k-anonymous or when the quality of the solution is worse.

Influence of the parameters on the algorithm

- with the increase of k, the level of **privacy goes up** but other **metrics go down**
- **execution time** also **increases** with higher k values

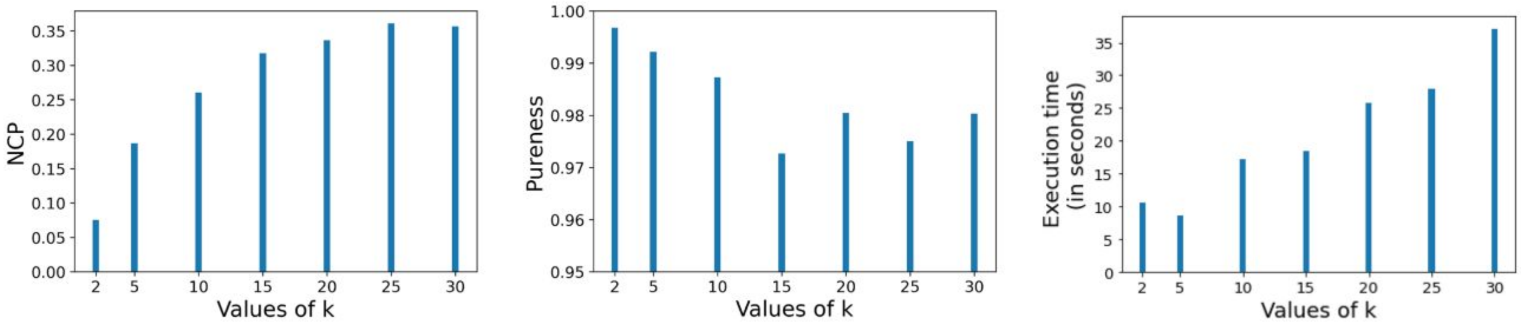

Figure: Influence of the k parameter on different metrics of the solution. Source: original paper.

Influence of the parameters on the algorithm

- with the increase of k, the level of **privacy goes up** but other **metrics go down**
- **execution time** also **increases** with higher k values

- increasing α i**ncreases NCP** but **lowers** the **pureness**
- **execution time decreases** with higher α values

Influence of the parameters on the algorithm

- with the increase of k, the level of **privacy goes up** but other **metrics go down**
- **execution time** also **increases** with higher k values

- increasing α i**ncreases NCP** but **lowers** the **pureness**
- **execution time decreases** with higher α values

- increasing the **number of iterations improves** all **metrics** but increases execution time

Evaluation

Table 5. Description of Used Datasets with Dataset Properties

Dataset	Adult[7]	CMC[8]	German[9]	Heart[10]	Hospital[11]	Informs[12]
# instances	48,842	1,473	1,000	303	8,160	5,000
# attributes	11	8	19	12	20	13
QID	*Age, Sex, Race, Relationship, Marital status*	*WifeAge, ChildrenBorn*	*Age, Foreign Personal status, Residence time, Employment, Job, Property, Housing*	*Age, Sex*	*Age Group, Race, Gender, Ethnicity, Zip Code*	*Dobmm, Dobyy, Sex, Marry, Educyear*
Sensitive attribute	*Sex*	*WifeReligion*	*Personal status*	*Sex*	*Gender*	*Race*
Target attribute	*Income*	*Contraceptive method*	*Credit decision*	*Heart disease*	*Costs*	*Income*
Uniquely identifiable (in %)	3.17	4.41	83.7	4.62	6.32	76.18
$\lvert EQ \rvert < 10$ (in %)	15.39	53.78	100	79.54	37.08	100

Explanation of the |EQ| < 10 row: "We measure the percentage of instances that are not protected by k-anonymity (with k = 10). This thus means that we measure the percentage of people that belong to an equivalence class with a size smaller than 10." **- quote from the source article**

Making the whole dataset k-anonymous

- k-anonymity is an existing property when considering whole datasets.

- Making a whole dataset k-anonymous means, that if **any** observation is released, it should be indistinguishable between k-1 other observations.

This is a much stronger property than counterfactual k-anonymity

Results

Table 6. Results of CF-K Over All the Datasets ($k = 10$)

Dataset	Adult	CMC	German	Heart	Hospital	Informs
NCP (mean)	0.55%	3.84%	21.41%	2.81%	3.42%	9.97%
Pureness (mean)	99.81%	93.15%	98.52%	100%	91.39%	85.33%
Execution time (mean)	24.78s	16.20s	13.31s	3.93s	17.76s	32.20s
C_{DM}	87,181	5,366	1,010	790	17,115	9,023
$\frac{C_{DM}}{\#explanations}$	110.78	13.2	16.83	14.11	22.94	13.65
CM	0.82	0.28	0.03	0.32	0.77	0.12

Table 7. Results of the Mondrian Algorithm ($k = 10$)

Dataset	Adult	CMC	German	Heart	Hospital	Informs
NCP (mean)	15.97%	7.05%	59.55%	53.01%	26.03%	36.31%
Pureness (mean)	90.30%	69.15%	90.50%	100%	63.77%	72.40%
Execution time (mean)	7.11s	0.87s	0.38s	0.23s	1.19s	1.11s
C_{DM} **(mean)**	120,227	6,318	963	1,044	16,534	9,177
$\frac{C_{DM}}{\#explanations}$	152.77	15.56	16.05	18.64	22.16	13.88
CM (mean)	0.83	0.24	0.17	0.41	0.80	0.40

Results

Table 8. Plausibility Results for Various Settings of the NICE Algorithm and CF-K, Lower Values are Better (Closer to the Data Manifold)

	NICE (none)	NICE (sparse)	NICE (prox)	NICE (plaus)	CF-K (k = 5)	CF-K (k = 10)	CF-K (k = 20)
1NN	0	2.77	2.94	2.48	0.84	**1.22**	1.32
5NN	2.64	3.73	3.81	3.54	2.72	2.80	2.83

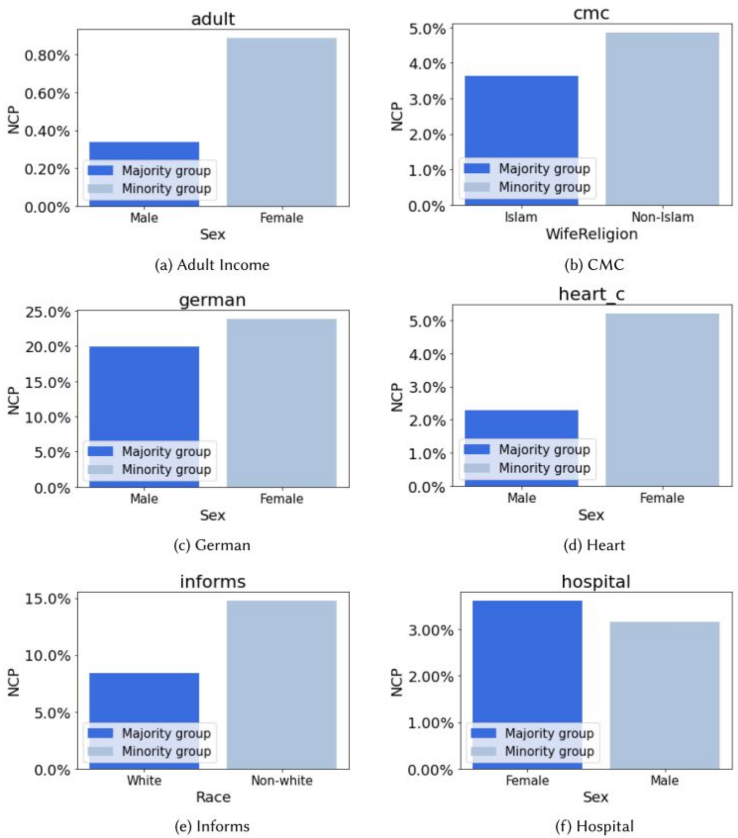

Figure: Comparison of the NCP metric in the minority and majority group. Source: original paper.

Thanks for the attention!

Questions and discussion

Mikołaj Spytek

MI2.AI Seminar, Warsaw, 15th April 2024