Consiglio Nazionale delle Ricerche
Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni

Counterfactual explanations through Support Vector Data Description

Methodology and examples

Methodology

- Suppose we have a dataset $\mathcal{X} \times \mathcal{Y} \subset R^N \times \{-1, +1\}, N \geq 2$ consisting of controllable features $\boldsymbol{u}$ and non-controllable features $\boldsymbol{z}$.

  An observation $\boldsymbol{x} \in \mathcal{X}$ can be described as:

$$\boldsymbol{x} = \left(u^1, u^2, \ldots, u^n, z^1, z^2, \ldots, z^m\right) \in R^{n+m=N}$$

- Considering a binary classification problem, we perform a TC-SVDD classification obtaining two regions:

$$S_1 \doteq \{\boldsymbol{x} \in R^N : \|\boldsymbol{x} - \mathbf{a}_1\|^2 \leq R_1^2, \|\boldsymbol{x} - \mathbf{a}_2\|^2 \geq R_2^2\}$$

$$S_2 \doteq \{\boldsymbol{x} \in R^N : \|\boldsymbol{x} - \mathbf{a}_2\|^2 \leq R_2^2, \|\boldsymbol{x} - \mathbf{a}_1\|^2 \geq R_1^2\}$$

Methodology

❑ **Counterfactual search:**

Given $\boldsymbol{x} = (\mathbf{u}, \mathbf{z}) \in S_1$ we want to determine the minimum variation of controllable variables only $\Delta\mathbf{u}^*$, so that $\boldsymbol{x}^* = (\mathbf{u} + \Delta\mathbf{u}^*, \mathbf{z})$ and $\boldsymbol{x}^* \in S_2$

$\Delta\mathbf{u}^*$ can be found by solving the following minimization problem:

$$
\begin{aligned}
\min_{\Delta\mathbf{u}\in\mathbb{R}^n} \quad & d(\boldsymbol{x}, (\mathbf{u} + \Delta\mathbf{u}, \mathbf{z})) \\
\text{subject to} \quad & \|(\mathbf{u} + \Delta\mathbf{u}, \mathbf{z}) - \mathbf{a}_2\|^2 \leq R_2^2 \\
& \|(\mathbf{u} + \Delta\mathbf{u}, \mathbf{z}) - \mathbf{a}_1\|^2 \geq R_1^2
\end{aligned}
$$

Numerical solution

Algorithm 1 `CounterfactualSVDD`

Dataset $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^N \times \{-1, +1\}$ is divided in training set $\mathcal{X}_{tr} \times \mathcal{Y}_{tr}$ and validation set $\mathcal{X}_{vl} \times \mathcal{Y}_{vl}$.

A TC-SVDD [12] is performed on $\mathcal{X}_{tr} \times \mathcal{Y}_{tr}$ and validated on $\mathcal{X}_{vl} \times \mathcal{Y}_{vl}$ in order to derive S_1 and S_2.

$N_C > 0$ is fixed.

1. $\mathcal{C} = [\,]$
2. **Sample** quasi-randomly a new dataset G
3. $G_1 \cup G_2 \doteq G \cap (S_1 \triangle S_2)$
4. **for** $i = 1 : N_C$
4.1 $\boldsymbol{x}_i = (\mathbf{u}_i, \mathbf{z}_i) \in S_1$
4.2 $d_i = d\left(\boldsymbol{x}_i, G_{2_{|\mathbf{z}=\mathbf{z}_i}}\right)$
4.3 $\boldsymbol{x}'_i = \min(d_i)$
4.4 **if**$(\boldsymbol{x}_i \in S_1 \ \& \ \boldsymbol{x}'_i \in S_2)$
4.4.1 $\mathcal{C} = \mathcal{C} \cup \{\boldsymbol{x}'_i\}$
4.5 **end**
5. **end**
6. **return** $\mathcal{C}$

Set of counterfactuals points belonging to S_2

Line	Symbol	Description
1.	$\mathcal{C}$	Set of counterfactuals
3.	$\triangle$	Symmetric difference, $G_1 = G \cap (S_1 \setminus S_2)$, $G_2 = G \cup (S_2 \setminus S_1)$
4.1	$\boldsymbol{x}_i$	Factual point
4.2	d	Distance function
4.2	$G_{2_{\mathbf{z}=\mathbf{z}_i}}$	G_2 points with component $\mathbf{z}$ equal to $\mathbf{z}_i$
4.4.1	$\boldsymbol{x}'_i$	Counterfactual point

Methodology

- **Computational cost**
  - SVDD:

$$O(SVDD) = \mathrm{O}(\max(\mathrm{n}, \mathrm{d})\min(n, d)^2)$$

  - Counterfactual search:

$$O(SC) = \mathrm{O}(\max(\mathrm{q}, N_C \max(\mathrm{D}, \mathrm{g})))$$

- **Counterfactual distance (CD)**

$$\mathrm{CD} = \mathrm{d}(\boldsymbol{a}_1, \boldsymbol{x}') - R_1$$

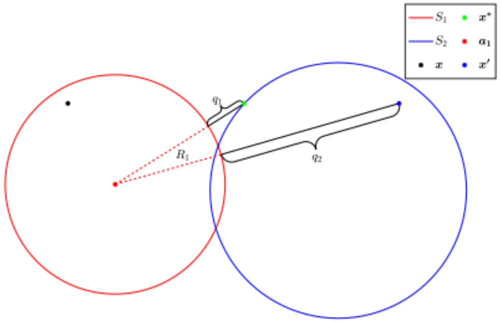
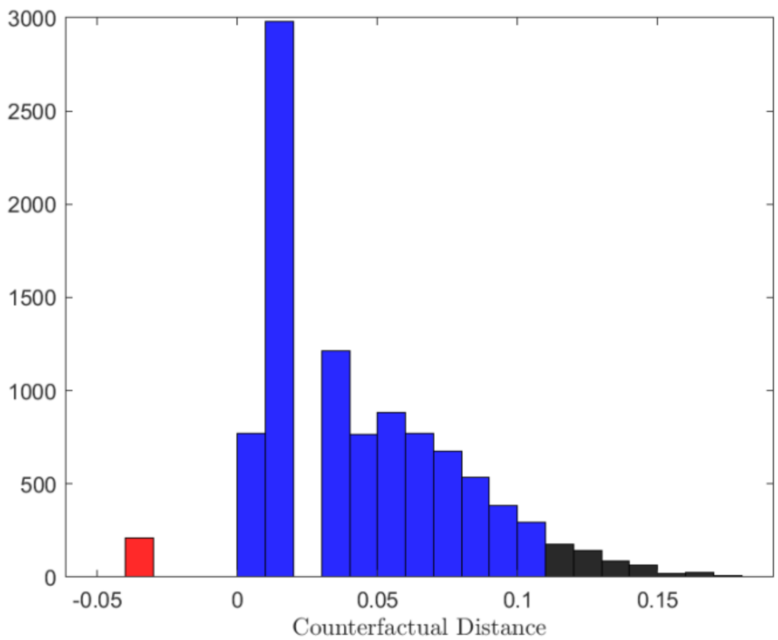


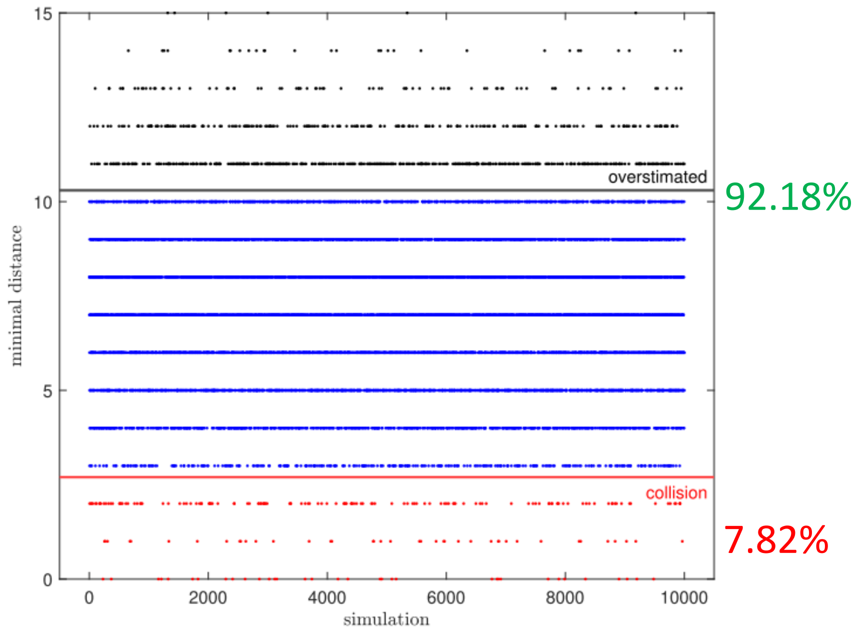
FIGURE 2. 2D-linear example of CD: this metric evaluates the goodness of the counterfactual, the closer q is to zero the more the counterfactual is optimal in terms of minimum distance. In the figure, $q_2 > q_1$ and the blue counterfactual x' is worst than the green (optimal) one x^*.

Example 1: truck platooning

OBJECTIVE: avoid collision in a platoon by acting on the mutual distance and speed between each pair of vehicles in the initial condition (t=0)



(a) CD of extracted counterfactuals. The red bin refers to counterfactuals that are incorrect, i.e. $q < 0$. Black bins refer to counterfactuals that overestimate corrections ($q > 0.1$).



(b) Behaviour of simulations with counterfactuals extracted via **Algorithm 1**. The platoon collides when the minimum distance in the simulation is less than or equal to 2 (red dots). Black dots refer to counterfactuals that overestimate the correction (minimum distance greater than 10).

Example 2: disease prevention

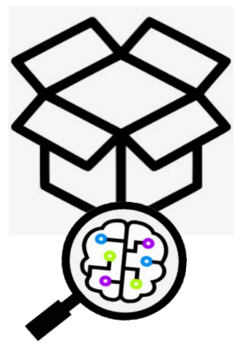
OBJECTIVE: find minimum yet significant changes in biomarker values that allow to reduce the risk of developing diabetes <u>on an individual basis</u>

Input features

- Age
- Gender
- Hypertension
- sBP
- BMI
- LDL
- HDL
- TG
- FBS

Output

- highT2DM
- lowT2DM



Counterfactual Explanations
+
FNR control

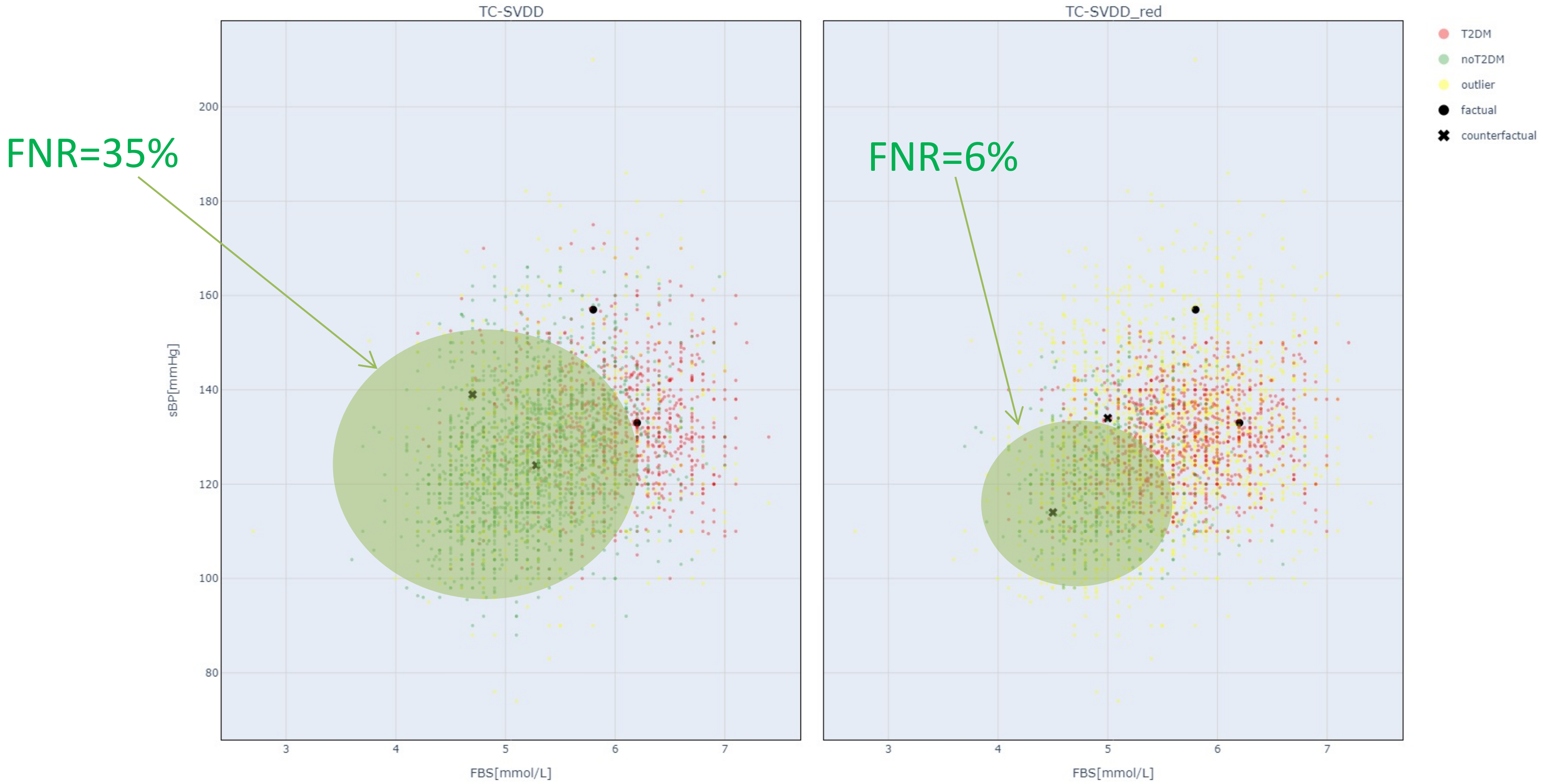
Increased transparency

Application-grounded evaluation

- **General questions on AI**
- **T2DM Risk evaluation**
- **Counterfactuals evaluation**

highT2DM → lowT2DM

Example 2: TC-SVDD with FNR reduction



Example 2: Expert assessment

- <u>Example of question:</u>

A *female* patient with the following biomarkers is at high risk of developing T2DM (1 year estimation):

Gender	Age	FBS $[mmol/L]$	BMI $[kg/m^2]$	sBP $[mmHg]$	LDL $[mmol/L]$	HDL $[mmol/L]$	TG $[mmol/L]$	Total Chol $[mmol/L]$	HTN
Female	63	6.2	28.7	133	3.1	1.1	1.5	4.9	Yes

The algorithm proposes to lower the risk of developing T2DM by suggesting a strategy that targets the following values:

FBS $[mmol/L]$	BMI $[kg/m^2]$	sBP $[mmHg]$	LDL $[mmol/L]$	HDL $[mmol/L]$	TG $[mmol/L]$	Total Chol $[mmol/L]$	HTN
4.5	25	114	3.0	0.8	0.4	3.8	

How much do you agree with the algorithm proposal?

moderately agree (5), strongly agree (2)

→ **Appropriate targets**

→ **Proposed treatment:** lifestyle changes including moderate healthy diet and regular exercise

Example 3: crowding prediction (subway station)

OBJECTIVE: find minimum yet significant changes in modifiable parameters to prevent possible crowding situations in a near future (e.g., 15 minutes time window)

Factual

Users arriving by train from Brin (T-15min) = 63
&
Users entering deFerrari station (T-30min) = 55
&
Trains arrival frequency (T-15min) = 15 min

Crowded station at time T

Explainable AI

Counterfactual Explanation

Users arriving by train from Brin (T-15min) = 42
&
Users entering deFerrari station (T-30min) = 25
&
Trains arrival frequency (T-15min) = 11 min

Not-crowded station at time T

?

Expert assessment

Example 3: crowding prediction (subway station)

EXPERT VALIDATION: Assess the feasibility of recommendations generated by an interpretable artificial intelligence model based on simulated data.

- **Completion time:** 15–20 minutes
- **Base structure:**
  - Collection of generic information, such as degree of AI knowledge, trust in AI, domain expertise …
  - Specific evaluation of a set of suggested counterfactuals-based recommendations for the purpose of crowding prevention (*5-items Likert scale*)
  - General evaluation of the method in terms of *realism* and *applicability* of the proposed recommendations
  - Request any additional variables to be considered in the simulation

Example 3: crowding prediction (subway station)

Scenario F is characterized by a number of people on the quay in the Brignole direction exceeding 30. The AI algorithm suggests to avoid exceeding the threshold by changing the variables as proposed in *C (all parameters are free to change), C VT (train-related parameters are constrained) or C VP (people-related parameters are constrained)*.

How much do you agree with the suggestions proposed by the algorithm?

V1: 'Incoming users T-30Min'
V2: 'Incoming users T-15Min'
V3: 'Trains arrival frequency T-30Min'
V4: 'Trains arrival frequency T-15Min'
V5: 'Train occupancy T-15Min'
V6: 'Users waiting for the train T-15Min'
V7: ' Maximum number of People (stairs) T-15Min'