Counterfactual Explanations Under Learning From Interpretation Transition

Tony Ribeiro[1,3,4], Maxime Folschette[2], Morgan Magnin[1], Katsumi Inoue[3], Kotaro Okazaki[4], Kuo-Yen Lo[4], Tuan Nguyen[5], Jérémie Poschmann[6], Antoine Roquilly[6,7]

1. Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France
2. Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France.
3. National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan.
4. Steelous Protocol, 8-20-32, Ginza, Chuo-ku, Tokyo 104-0061, Japan.
5. Hanoi University of Science and Technology, No. 1 Dai Co Viet, Hai Ba Trung, Ha Noi, Vietnam.
6. Inserm, Nantes Université, CHU Nantes, Center for Research in Transplantation and Translational Immunology, UMR 1064, ITUN2, F-44000, Nantes, France.
7. Nantes Université, CHU Nantes, INSERM, Anesthesie Réanimation, F-44000 Nantes, France.

9th October 2025, Nantes, France

Outline

Outline

1. Motivations
2. LFIT Overview
3. Counterfactual under LFIT
   - CELOS Algorithm
4. Conclusions

HOMI-LUNG: A strong consortium to fill the gap.
Pneumonia is a heart matter too (and vice-versa).

HOMI-LUNG

Funded by
the European Union

"Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or ERCEA. Neither the European Union nor the granting authority can be held responsible for them."

Motivations: machine-produced Decision Making

Algorithms used for decision-making impact human lives in many domains.

- Credit lending
- Talent sourcing
- Medical treatment

Rising interests regarding fairness, accountability and transparency.

Motivations: eXplainable Artificial Intelligence (XAI)

Explanations for machine-produced decisions has several motivations.

Organization benefit

- Useful to check for bias and unsure fairness.
- Can help the models developers identify performance issues.
- Help adhere to laws for machine-produced decisions, e.g., GDPR

User benefits:

- Personalized feedback highlights key factors influencing outcome.
- Empowering actionable insights to lead to favorable outcomes.
- Various explanations serve transparency and trustworthiness.

Counterfactual Explanations

Counterfactuals involve considering alternatives that affect decisions.

Example

Bank denied a loan application because: "credit score is insufficient".

- It does not give insight to the customer to change the outcome.

Counterfactual explanations:

- "Annual income is 30,000$, changing to 45,000$ makes loan ok"

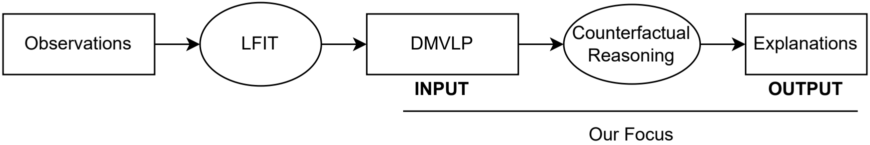
In this example we get two kind of insight from counterfactuals:

- Decision causes: Income below threshold led to denial.
- Potential solution: increase income by 15,000$ for approval.

It allows to comprehend actions taken and avoid similar issues in future.

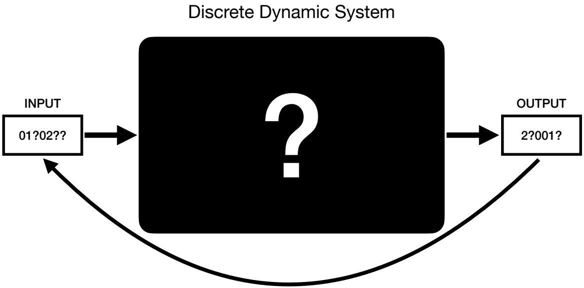
Counterfactual reasoning from LFIT

Goal: produce counterfactual explanations from LFIT output.



Outline

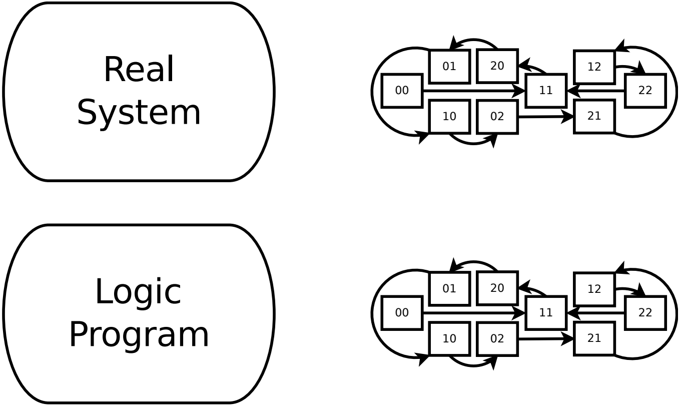
Learning From Interpretation Transitions



- **Idea:**
  Learn black-box **internal mechanics** from its input/output states.
- **Discrete System:**
  Input/output are vectors containing discrete and unknowns values.
- **Dynamic System:**
  System output become the next input (require vectors of same size).

Learning From Interpretation Transitions

Goal: produce a logic program with the same behavior as the system observed.



Logic Rule

$$\underbrace{\mathrm{v}_0^{val_0}}_{\substack{\textit{head}\\ \text{target atom}}} \leftarrow \underbrace{\mathrm{v}_1^{val_1} \wedge \mathrm{v}_2^{val_2} \wedge \ldots \wedge \mathrm{v}_n^{val_n}}_{\substack{\textit{body}\\ \text{feature atoms}}}.$$

Logic Rule

$$\underbrace{\mathrm{v}_0^{val_0}}_{\substack{\textit{head}\\ \text{target atom}}} \leftarrow \underbrace{\mathrm{v}_1^{val_1} \;\wedge\; \mathrm{v}_2^{val_2} \;\wedge\; \dots \;\wedge\; \mathrm{v}_n^{val_n}}_{\substack{\textit{body}\\ \text{feature atoms}}}.$$

- v_0, v_1, v_2, ..., v_n: variables a_t, a_{t-1}, b_t, b_{t-1}, z_t, z_{t-1}
  - Variables are split into feature ($\mathcal{F}$) and target ($\mathcal{T}$) variables
  - $\mathrm{v}_0 \in \mathcal{T}$ a_t, b_t, z_t
  - $\mathrm{v}_1, \mathrm{v}_2, \dots, \mathrm{v}_n \in \mathcal{F}$ a_{t-1}, b_{t-1}, z_{t-1}
  - Implicit time step: t in the *head* and $t-1$ in the *body*

Logic Rule

$$\underbrace{\mathrm{v}_0^{val_0}}_{\substack{\textit{head} \\ \text{target atom}}} \leftarrow \underbrace{\mathrm{v}_1^{val_1} \wedge \mathrm{v}_2^{val_2} \wedge \ldots \wedge \mathrm{v}_n^{val_n}}_{\substack{\textit{body} \\ \text{feature atoms}}}.$$

- v_0, v_1, v_2, ..., v_n: variables $a_t,\ a_{t-1},\ b_t,\ b_{t-1},\ z_t,\ z_{t-1}$
  - Variables are split into feature ($\mathcal{F}$) and target ($\mathcal{T}$) variables
  - $\mathrm{v}_0 \in \mathcal{T}$ $a_t,\ b_t,\ z_t$
  - $\mathrm{v}_1, \mathrm{v}_2, \ldots, \mathrm{v}_n \in \mathcal{F}$ $a_{t-1},\ b_{t-1},\ z_{t-1}$
  - Implicit time step: t in the *head* and $t-1$ in the *body*
- val_0, val_1, val_2, ..., val_n: values 0, 1, 2, ...
  - $val_i \in \mathrm{dom}(\mathrm{v}_i)$

Logic Rule

$$\underbrace{\mathrm{v}_0^{val_0}}_{\substack{\textit{head}\\ \text{target atom}}} \leftarrow \underbrace{\mathrm{v}_1^{val_1} \wedge \mathrm{v}_2^{val_2} \wedge \ldots \wedge \mathrm{v}_n^{val_n}}_{\substack{\textit{body}\\ \text{feature atoms}}}.$$

- $\mathrm{v}_0, \mathrm{v}_1, \mathrm{v}_2, \ldots, \mathrm{v}_n$: variables $\quad a_t,\ a_{t-1},\ b_t,\ b_{t-1},\ z_t,\ z_{t-1}$
  - Variables are split into feature ($\mathcal{F}$) and target ($\mathcal{T}$) variables
  - $\mathrm{v}_0 \in \mathcal{T}$ $\quad a_t,\ b_t,\ z_t$
  - $\mathrm{v}_1, \mathrm{v}_2, \ldots, \mathrm{v}_n \in \mathcal{F}$ $\quad a_{t-1},\ b_{t-1},\ z_{t-1}$
  - Implicit time step: t in the *head* and $t-1$ in the *body*
- $val_0, val_1, val_2, \ldots, val_n$: values $\quad 0, 1, 2, \ldots$
  - $val_i \in \mathrm{dom}(\mathrm{v}_i)$
- All atoms in the *body* are in conjunction
- $\leftarrow$ is the (reverse) implication

Interpretation of a Logic Rule

$$\underbrace{\mathrm{v}_0^{val_0}}_{\substack{\textit{head}\\ \text{target atom}}} \leftarrow \underbrace{\mathrm{v}_1^{val_1} \wedge \mathrm{v}_2^{val_2} \wedge \ldots \wedge \mathrm{v}_n^{val_n}}_{\substack{\textit{body}\\ \text{feature atoms}}}.$$

Interpretation of a Logic Rule

$$\underbrace{\mathrm{v}_0^{val_0}}_{\substack{head \\ \text{target atom}}} \leftarrow \underbrace{\mathrm{v}_1^{val_1} \wedge \mathrm{v}_2^{val_2} \wedge \ldots \wedge \mathrm{v}_n^{val_n}}_{\substack{body \\ \text{feature atoms}}}.$$

Interpretation: When *body* is true, *head* is a potential outcome

Interpretation of a Logic Rule

$$\underbrace{\mathrm{v}_0^{val_0}}_{\substack{\textit{head}\\ \text{target atom}}} \leftarrow \underbrace{\mathrm{v}_1^{val_1} \ \wedge\ \mathrm{v}_2^{val_2} \ \wedge\ \dots\ \wedge\ \mathrm{v}_n^{val_n}}_{\substack{\textit{body}\\ \text{feature atoms}}}.$$

Interpretation: When *body* is true, *head* is a potential outcome

Examples:

$$a_t^1 \leftarrow a_{t-1}^2 \wedge b_{t-1}^0 \wedge z_{t-1}^1.$$
$$b_t^1 \leftarrow z_{t-1}^1.$$
$$z_t^0 \leftarrow \top.$$

Interpretation of a Logic Rule

$$\underbrace{\mathrm{v}_0^{val_0}}_{\substack{head \\ \text{target atom}}} \leftarrow \underbrace{\mathrm{v}_1^{val_1} \;\wedge\; \mathrm{v}_2^{val_2} \;\wedge\; \dots \;\wedge\; \mathrm{v}_n^{val_n}}_{\substack{body \\ \text{feature atoms}}}.$$

Interpretation: When *body* is true, *head* is a potential outcome

$$\text{Examples:} \quad \left.\begin{array}{l} a_t^1 \leftarrow a_{t-1}^2 \wedge b_{t-1}^0 \wedge z_{t-1}^1. \\ b_t^1 \leftarrow z_{t-1}^1. \\ z_t^0 \leftarrow \top. \end{array}\right\} \text{ all } \textbf{match } \langle a_{t-1}^2, b_{t-1}^0, z_{t-1}^1 \rangle$$

A rule R **matches** a state s iff $body \subseteq s$

Interpretation of a Logic Rule

$$\underbrace{v_0^{val_0}}_{\substack{\textit{head}\\ \text{target atom}}} \leftarrow \underbrace{v_1^{val_1} \ \wedge \ v_2^{val_2} \ \wedge \ \ldots \ \wedge \ v_n^{val_n}}_{\substack{\textit{body}\\ \text{feature atoms}}}.$$

Interpretation: When *body* is true, *head* is a potential outcome

$$\text{Examples:} \quad \left.\begin{array}{l} a_t^1 \leftarrow a_{t-1}^2 \wedge b_{t-1}^0 \wedge z_{t-1}^1. \\ b_t^1 \leftarrow z_{t-1}^1. \\ z_t^0 \leftarrow \top. \end{array}\right\} \text{ all } \textbf{match } \langle a_{t-1}^2, b_{t-1}^0, z_{t-1}^1 \rangle$$

A rule R **matches** a state s iff *body* $\subseteq s$

Interpretation: When a rule **matches** a state,
the rule's *head* becomes a **candidate** for the next state

Interpretation of a Logic Rule

$$\underbrace{\mathrm{v}_0^{val_0}}_{\substack{\textit{head}\\ \text{target atom}}} \leftarrow \underbrace{\mathrm{v}_1^{val_1} \;\wedge\; \mathrm{v}_2^{val_2} \;\wedge\; \dots \;\wedge\; \mathrm{v}_n^{val_n}}_{\substack{\textit{body}\\ \text{feature atoms}}}.$$

Interpretation: When *body* is true, *head* is a potential outcome

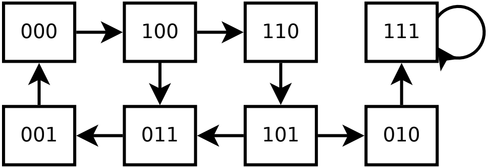
$$\text{Examples:}\quad \left.\begin{array}{l} a_t^1 \leftarrow a_{t-1}^2 \wedge b_{t-1}^0 \wedge z_{t-1}^1. \\ b_t^1 \leftarrow z_{t-1}^1. \\ z_t^0 \leftarrow \top. \end{array}\right\} \text{all } \textbf{match } \langle a_{t-1}^2, b_{t-1}^0, z_{t-1}^1 \rangle$$

A rule R **matches** a state s iff $body \subseteq s$

Interpretation: When a rule **matches** a state,
the rule's *head* becomes a **candidate** for the next state

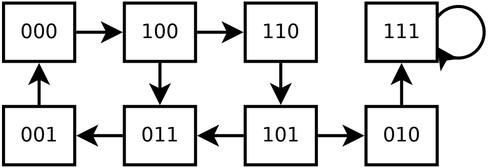
Semantics = From this information, what are the next possible state(s)?
(Similar to discrete networks)

Example: LFIT with complete observations



Features: $\{a_{t-1}, b_{t-1}, c_{t-1}\}$, Targets: $\{a_t, b_t, c_t\}$, Domains: $\{0, 1\}$

Example: LFIT with complete observations



Features: $\{a_{t-1}, b_{t-1}, c_{t-1}\}$, Targets: $\{a_t, b_t, c_t\}$, Domains: $\{0, 1\}$

$a_t^1 \leftarrow c_{t-1}^0$
$a_t^1 \leftarrow a_{t-1}^1, b_{t-1}^1$

$a_t^0 \leftarrow a_{t-1}^0, c_{t-1}^1$
$a_t^0 \leftarrow a_{t-1}^1, b_{t-1}^0$
$a_t^0 \leftarrow b_{t-1}^0, c_{t-1}^1$

$b_t^1 \leftarrow a_{t-1}^1, b_{t-1}^0$
$b_t^1 \leftarrow a_{t-1}^1, c_{t-1}^1$
$b_t^1 \leftarrow a_{t-1}^0, b_{t-1}^1, c_{t-1}^0$

$b_t^0 \leftarrow a_{t-1}^0, b_{t-1}^0$
$b_t^0 \leftarrow a_{t-1}^0, c_{t-1}^1$
$b_t^0 \leftarrow a_{t-1}^1, b_{t-1}^1, c_{t-1}^0$

$c_t^1 \leftarrow a_{t-1}^1$
$c_t^1 \leftarrow b_{t-1}^1$

$c_t^0 \leftarrow b_{t-1}^0$

Outline

Problem Definition

Definition (Counterfactual Explanation Problem)

Let P be a $\mathcal{DM}\mathrm{VLP}$, $s \in \mathcal{S}^{\mathcal{F}}$ be a feature state, $\mathrm{v} \in \mathcal{T}$ be a target variable. A counterfactual explanation problem is a tuple $CP := (P, s, \mathrm{v}, Val_{out}, Val_{in})$ where $Val_{out} \subset \mathrm{dom}(\mathrm{v})$ and $Val_{in} \subset \mathrm{dom}(\mathrm{v})$ such that $Val_{out} \cap Val_{in} = \emptyset$. A solution to CP is a set of atoms $X = s' \setminus s$ for some feature state s' such that:

- no rule R of P with $head(R) = \mathrm{v}^{val}, val \in Val_{out}$ matches s' and
- there is a rule R' of P with $head(R') = \mathrm{v}^{val'}, val' \in Val_{in}$ that matches s'.

A solution to CP is minimal if no subset of the solution is also a solution.

Counterfactual Explanation Example

Example

Consider the following $\mathcal{DM}\mathrm{VLP}$ P such that $\mathcal{F} = \{a, b, c\}, \mathcal{T} = \{y\}, \mathrm{dom}(a) = \mathrm{dom}(b) = \{0, 1\}, \mathrm{dom}(c) = \mathrm{dom}(y) = \{0, 1, 2\}$:

$$
\begin{array}{lll}
y^0 \leftarrow a^0 & y^1 \leftarrow b^1 & y^2 \leftarrow a^1 \wedge b^1 \wedge c^2 \\
y^0 \leftarrow b^0 & y^1 \leftarrow a^0 \wedge c^1 & \\
y^0 \leftarrow c^0 & y^1 \leftarrow c^2 &
\end{array}
$$

Counterfactual Explanation Example

Example

Consider the following $\mathcal{DM}\mathrm{VLP}$ P such that $\mathcal{F} = \{a, b, c\}, \mathcal{T} = \{y\}, \mathrm{dom}(a) = \mathrm{dom}(b) = \{0, 1\}, \mathrm{dom}(c) = \mathrm{dom}(y) = \{0, 1, 2\}$:

$$
\begin{array}{lll}
y^0 \leftarrow a^0 & y^1 \leftarrow b^1 & y^2 \leftarrow a^1 \wedge b^1 \wedge c^2 \\
y^0 \leftarrow b^0 & y^1 \leftarrow a^0 \wedge c^1 & \\
y^0 \leftarrow c^0 & y^1 \leftarrow c^2 &
\end{array}
$$

Let $CP := (P, s, y, \{y^1\}, \{y^0, y^2\})$ where $s = \{a^0, b^1, c^1\}$.

Counterfactual Explanation Example

Example

Consider the following $\mathcal{DM}\mathrm{VLP}$ P such that $\mathcal{F} = \{a, b, c\}, \mathcal{T} = \{y\}, \mathrm{dom}(a) = \mathrm{dom}(b) = \{0, 1\}, \mathrm{dom}(c) = \mathrm{dom}(y) = \{0, 1, 2\}$:

$y^0 \leftarrow a^0$	$y^1 \leftarrow b^1$	$y^2 \leftarrow a^1 \wedge b^1 \wedge c^2$
$y^0 \leftarrow b^0$	$y^1 \leftarrow a^0 \wedge c^1$	
$y^0 \leftarrow c^0$	$y^1 \leftarrow c^2$	

Let $CP := (P, s, y, \{y^1\}, \{y^0, y^2\})$ where $s = \{a^0, b^1, c^1\}$.

Counterfactual Explanation Example

Example

Consider the following $\mathcal{DM}\mathrm{VLP}$ P such that $\mathcal{F} = \{a, b, c\}, \mathcal{T} = \{y\}, \mathrm{dom}(a) = \mathrm{dom}(b) = \{0, 1\}, \mathrm{dom}(c) = \mathrm{dom}(y) = \{0, 1, 2\}$:

$$
\begin{array}{lll}
y^0 \leftarrow a^0 & y^1 \leftarrow b^1 & y^2 \leftarrow a^1 \wedge b^1 \wedge c^2 \\
y^0 \leftarrow b^0 & y^1 \leftarrow a^0 \wedge c^1 & \\
y^0 \leftarrow c^0 & y^1 \leftarrow c^2 &
\end{array}
$$

Let $CP := (P, s, y, \{y^1\}, \{y^0, y^2\})$ where $s = \{a^0, b^1, c^1\}$.

- The minimal solutions for y^0 are $\{\{a^1, b^0\}, \{b^0, c^0\}\}$.

Counterfactual Explanation Example

Example

Consider the following $\mathcal{DM}\mathrm{VLP}$ P such that $\mathcal{F} = \{a, b, c\}, \mathcal{T} = \{y\}$, $\mathrm{dom}(a) = \mathrm{dom}(b) = \{0, 1\}, \mathrm{dom}(c) = \mathrm{dom}(y) = \{0, 1, 2\}$:

$$
\begin{array}{lll}
y^0 \leftarrow a^0 & y^1 \leftarrow b^1 & y^2 \leftarrow a^1 \wedge b^1 \wedge c^2 \\
y^0 \leftarrow b^0 & y^1 \leftarrow a^0 \wedge c^1 & \\
y^0 \leftarrow c^0 & y^1 \leftarrow c^2 &
\end{array}
$$

Let $CP := (P, s, y, \{y^1\}, \{y^0, y^2\})$ where $s = \{a^0, b^1, c^1\}$.

- The minimal solutions for y^0 are $\{\{a^1, b^0\}, \{b^0, c^0\}\}$.

Counterfactual Explanation Example

Example

Consider the following $\mathcal{DM}\mathrm{VLP}$ P such that $\mathcal{F} = \{a, b, c\}, \mathcal{T} = \{y\}, \mathrm{dom}(a) = \mathrm{dom}(b) = \{0, 1\}, \mathrm{dom}(c) = \mathrm{dom}(y) = \{0, 1, 2\}$:

$y^0 \leftarrow a^0$	$y^1 \leftarrow b^1$	$y^2 \leftarrow a^1 \wedge b^1 \wedge c^2$
$y^0 \leftarrow b^0$	$y^1 \leftarrow a^0 \wedge c^1$	
$y^0 \leftarrow c^0$	$y^1 \leftarrow c^2$	

Let $CP := (P, s, y, \{y^1\}, \{y^0, y^2\})$ where $s = \{a^0, b^1, c^1\}$.

- The minimal solutions for y^0 are $\{\{a^1, b^0\}, \{b^0, c^0\}\}$.
- No solution for y^2.

Naïve enumeration approach

- **INPUT:** $CP := (P, s, \mathrm{v}, Val_{out}, Val_{in})$.
- $out\_rules := \{R \in P \mid head(R) = \mathrm{v}^{val}, val \in Val_{out}\}$
- For each val of $Val_{in}, in\_rules_{val} := \{R \in P \mid head(R) = \mathrm{v}^{val}\}$
- For each val of $Val_{in}, solutions_{val} := \emptyset$
- For each complete state $s' \in \mathcal{S}^{\mathcal{F}}$:
  - If a rule of out_rules matches s': // Discard invalid state
    - ★ Continue
  - For each val of Val_{in}: // Extract valid changes to get v^{val}
    - ★ If a rule of in_rules_{val} matches s':
      $solutions_{val} := solutions_{val} \cup \{\{x \in s' \mid x \notin s\}\}$
- For each val of Val_{in}: // Keep only minimal sets
  - $solutions_{val} := \{x \in solutions_{val} \mid \nexists x' \in solutions_{val}, x' \subset x\}$
- **OUTPUT:** $\{solutions_{val} \mid val \in Val_{in}\}$

Outline

1. Motivations
2. LFIT Overview
3. Counterfactual under LFIT
    - CELOS Algorithm
4. Conclusions

Cross-matching

Definition (Rule cross-matching)

Let R, R' be two $\mathcal{M}\mathrm{VL}$ rules. The rules cross-match, if there exists a feature state $s \in \mathcal{S}^{\mathcal{F}}$ such that both R and R' match s.

Cross-matching

Definition (Rule cross-matching)

Let R, R' be two $\mathcal{M}\mathrm{VL}$ rules. The rules cross-match, if there exists a feature state $s \in \mathcal{S}^{\mathcal{F}}$ such that both R and R' match s.

Example

Let $\mathcal{A} = \{a, b, c, y\}$, domain of a, b is $\{0, 1\}$, domain of c, y is $\{0, 1, 2\}$:

- $y^1 \leftarrow a^0$, $y^1 \leftarrow b^1$ cross-match $\{a^0, b^1, c^0\}, \{a^0, b^1, c^1\}, \{a^0, b^1, c^2\}$.
- $y^1 \leftarrow b^1 \wedge c^0$ and $y^1 \leftarrow b^1$ cross-match $\{a^0, b^1, c^0\}, \{a^1, b^1, c^0\}$.
- $y^1 \leftarrow a^0 \wedge b^0$ and $y^1 \leftarrow b^1$ do not cross-match.

Cross-matching

Definition (Rule cross-matching)

Let R, R' be two $\mathcal{M}\mathrm{VL}$ rules. The rules cross-match, if there exists a feature state $s \in \mathcal{S}^{\mathcal{F}}$ such that both R and R' match s.

Example

Let $\mathcal{A} = \{a, b, c, y\}$, domain of a, b is $\{0, 1\}$, domain of c, y is $\{0, 1, 2\}$:

- $y^1 \leftarrow a^0$, $y^1 \leftarrow b^1$ cross-match $\{a^0, b^1, c^0\}, \{a^0, b^1, c^1\}, \{a^0, b^1, c^2\}$.
- $y^1 \leftarrow b^1 \wedge c^0$ and $y^1 \leftarrow b^1$ cross-match $\{a^0, b^1, c^0\}, \{a^1, b^1, c^0\}$.
- $y^1 \leftarrow a^0 \wedge b^0$ and $y^1 \leftarrow b^1$ do not cross-match.

Proposition

Let R, R' be two $\mathcal{M}\mathrm{VL}$ rules, they cross-match if and only if: for any v^{val} of $body(R)$, if there exists $\mathrm{v}^{val'} \in body(R')$, it implies that $val = val'$.

Anti-cross matching least specialization

Definition (Anti cross-matching least specialization)

Let R, R' be two $\mathcal{M}\mathrm{VL}$ rules that cross-match. The anti cross-matching least specialization of R by R' according to $\mathcal{A}$ is:

$$ACML_{\mathrm{spe}}(R, R', \mathcal{A}) := \{head(R) \leftarrow body(R) \cup \{\mathrm{v}^{val}\} \mid \mathrm{v}^{val'} \in body(R'), \mathrm{v}^{val} \in \mathcal{A}, val \neq val'\}.$$

Anti-cross matching least specialization

Definition (Anti cross-matching least specialization)

Let R, R' be two $\mathcal{M}\mathrm{VL}$ rules that cross-match. The anti cross-matching least specialization of R by R' according to $\mathcal{A}$ is:

$$ACML_{\mathrm{spe}}(R, R', \mathcal{A}) := \{head(R) \leftarrow body(R) \cup \{\mathrm{v}^{val}\} \mid \mathrm{v}^{val'} \in body(R'), \mathrm{v}^{val} \in \mathcal{A}, val \neq val'\}.$$

Proposition

Let R, R' be two $\mathcal{M}\mathrm{VL}$ rules. It holds that:

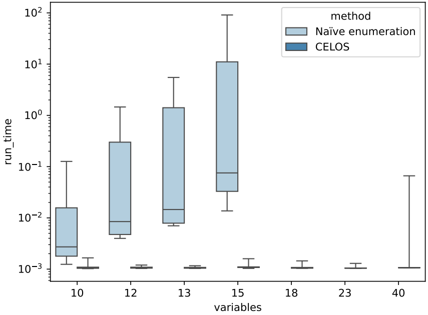
1. *For any R'' of $ACML_{\mathrm{spe}}(R, R', \mathcal{A})$, R'' and R' do not cross-match.*
2. *The feature states matched by $ACML_{\mathrm{spe}}(R, R', \mathcal{A})$ are all the states matched by R but not by R':*

$$\{s \in \mathcal{S}^{\mathcal{F}} \mid R'' \in ACML_{\mathrm{spe}}(R, R', \mathcal{A}), R'' \sqcap s\} = \{s \in \mathcal{S}^{\mathcal{F}} \mid R \sqcap s, R' \not\sqcap s\}.$$

CELOS

- **INPUT:** $CP := (P, s, \mathrm{v}, Val_{out}, Val_{in})$.
- $out\_rules := \{R \in P \mid head(R) = \mathrm{v}^{val}, val \in Val_{out}\}$
- For each val of Val_{in}, $in\_rules_{val} := \{R \in P \mid head(R) = \mathrm{v}^{val}\}$
- For each val of Val_{in}, $solutions_{val} := \emptyset$
- Initialize $P_{in} := \{\_ \leftarrow \emptyset\}$ // Rule head does not matter, only body is used
- For each rule R' of out_rules: // 1) Search necessary changes to avoid all Val_{out}
  - Extract and remove the rules of P_{in} that cross-match R':
    $M := \{R \in P_{in} \mid \forall \mathrm{v}^{val} \in body(R), \mathrm{v}^{val'} \in R' \implies val = val'\}$
  - $P_{in} := P_{in} \setminus M$
  - $LS := \emptyset$
  - For each rule R of M:
    - Compute its rule least specialization $P'_{in} = ACML_{\mathrm{spe}}(R, R', \mathcal{A})$
    - Remove rules in P'_{in} dominated by a rule in P_{in}
    - $LS := LS \cup P'_{in}$
  - Add all remaining rules of LS to P_{in}: $P_{in} := P_{in} \cup LS$
- For each val of Val_{in}: // 2) Compute necessary changes to produce a Val_{in}
  - For each R of in_rules_{val}, for each R' of in_rules_{val}:
    - If R, R' cross-match: // Combine and extract changes with state s
      $candidate := \{x \in (body(R) \cup body(R')) \mid x \notin s\}$
      $solutions_{val} := solutions_{val} \cup \{candidate\}$
  - $solutions_{val} := \{S \in solutions_{val} \mid \nexists S' \in solutions_{val}, S' \subset S\}$
- **OUTPUT:** $\{solutions_{val} \mid val \in Val_{in}\}$

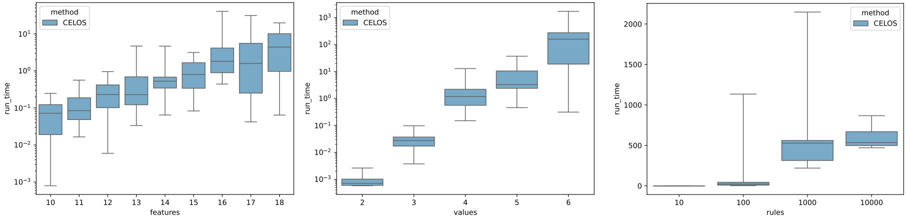
Evaluation



Runtime (in seconds, log scale) comparison of **CELOS** and **Naïve enumeration** on random counterfactual explanation problems.

- Boolean networks from Biological literature with 10 to 40 variables.
- 100 runs per target variable value, with a timeout of $1,000$ seconds.

Evaluation



Runtime (in seconds, log scale) of **CELOS** when applied to random counterfactual explanation problems from synthetic $\mathcal{DM}\mathrm{VLP}$ with a single target variable. The experiments used 10 runs for each problem setting.

- (1) Varying the number of variables from 10 to 18.
- (2) Increasing the domain size from 2 to 6.
- (3) Scaling the number of rules per target from 10 to $10,000$.

Outline

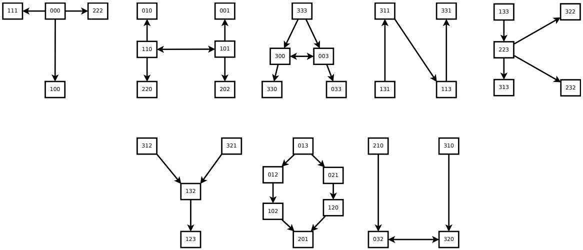
Conclusions

Contributions:

- Formalization of counterfactuals for dynamic multi-valued logic
- CELOS an efficient algorithm that leverage properties of DMVLP rules

Applications:

- Can be use to get more insight about the system learned by LFIT
- Analyze what can change a patient prognostic: decision robustness


111
000
222
100
010
001
110
101
220
202
333
300
003
330
033
311
331
131
113
133
322
223
313
232
312
321
132
123
013
012
021
102
120
201
210
310
032
320