Generating Contrastive Explanations for Inductive Logic Programming Based on a Near Miss Approach

Johannes Rabold, Michael Siebers, Ute Schmid

International Joint Conference on Learning and Reasoning

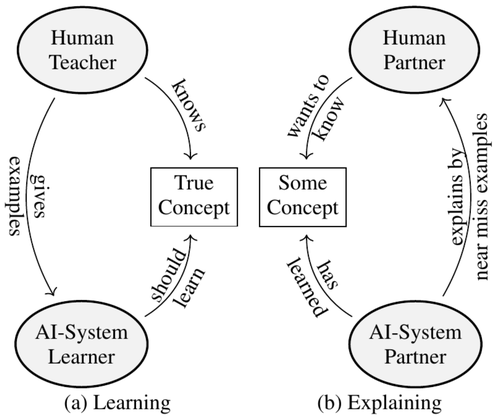
https://www.uni-bamberg.de/en/cogsys/

Introduction

Counterfactuals are explanations in the form of negative examples for a concept that only differ minimally from a given positive example.

You were denied a loan because your annual income was £30,000. If your income had been £45,000, you would have been offered a loan. (Wachter et al. 2017)

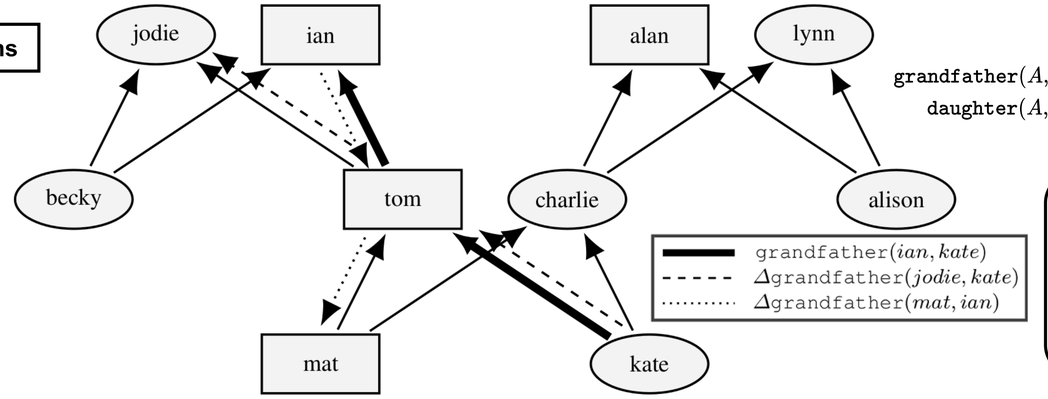
Such contrastive explanations were sucessfully applied in the image domain (e.g. CEM, Dhurandhar et al. 2018; MMD-critic, Kim et al. 2016) but to our knowledge not yet in relational domains.
Winston showed in 1970 that presenting near miss examples in a relational domain results in faster learning. We propose what is effective for learning is also effective for explaining a learned model. In this work, we present GeNME, an algorithm that generates helpful near miss explanations for relational models.



(a) Learning (b) Explaining

Domains

Family



$\texttt{grandfather}(A, B) \Longleftarrow \texttt{male}(A), \texttt{parent}(A, C), \texttt{parent}(C, B)$

$\texttt{daughter}(A, B) \Longleftarrow \texttt{female}(A), \texttt{child}(A, B)$

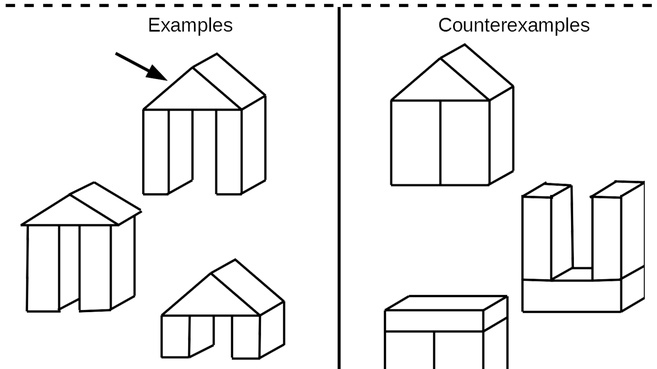
Predicates

$\texttt{female}(Person)$

$\texttt{male}(Person)$

$\texttt{parent}(Person1, Person2)$

Arches



$\texttt{arch}(A) \Longleftarrow \texttt{contains}(A, X), \texttt{contains}(A, Y), \texttt{contains}(A, Z),$
$\texttt{is\_a}(X, T), \texttt{is\_a}(Y, brick), \texttt{is\_a}(Z, brick),$
$\texttt{supports}(Y, X, A), \texttt{supports}(Z, X, A),$
$\texttt{not\_meets}(Y, Z, A)$

Predicates

$\texttt{is\_a}(Block, Type)$

$\texttt{contains}(Struct, Block)$

$\texttt{supports}(Block1, Block2, Struct)$

$\texttt{supported\_by}(Block1, Block2, Struct) \Longleftarrow \texttt{supports}(Block2, Block1, Struct)$

$\texttt{meets}(Block1, Block2, Struct)$

$\texttt{not\_meets}(Block1, Block2, Struct)$

GeNME: Generation of Near Miss Explanations

Require: Theory T, Finite set of rewriting filters O, Positive Example P

Initialize family of result sets $(\mathscr{E}_i)_{i\in\mathbb{N}}$ to empty sets
Initialize the set of all near miss candidates $\mathscr{N}$
for all local explanations $C\theta$ for P where $C \in T$ **do**
  for all $\mathscr{V}_{p\mapsto q} \in O$ **do**
    for all $\mathscr{L} \in \mathscr{V}_{p\mapsto q}(C)$ **do**
      create C' from C by replacing p with q in every literal in the body which is in $\mathscr{L}$
      for all $N \in \mathscr{N}$ **do**
        $d \leftarrow 0$
        $\mathscr{E} \leftarrow \{\}$
        while $\mathscr{E} = \{\}$ and $d < |\theta|$ **do**
          $d \leftarrow d + 1$
          for all partitions of θ into θ_1 and θ_2 such that $|\theta_2| = d$ **do**
            for all $\theta_2' = \{x_i \mapsto t_i' \mid x_i \mapsto t_i \in \theta_2\}$ such that no $t_i' = t_i$ **do**
              $E \leftarrow C'(\theta_1 \cup \theta_2')$
              $\mathscr{E} \leftarrow \mathscr{E} \cup \{E \mid T \models \text{body}(E) \text{ and head}(E) = N\}$
            end for
          end for
        end while
        $\mathscr{E}_d \leftarrow \mathscr{E}_d \cup \mathscr{E}$
      end for
    end for
  end for
end for
return $(\mathscr{E}_i)_{i\in\mathbb{N}}$

	gf(ian,kate) $\lvert\mathscr{N}\rvert = 96$	*dt(becky,jodie)* $\lvert\mathscr{N}\rvert = 92$
male ↔ female		
$\lvert\mathscr{E}_1\rvert$	1	1
$\lvert\mathscr{E}_2\rvert$	2	3
$\lvert\mathscr{E}_3\rvert$	1	0
parent ↔ child		
$\lvert\mathscr{E}_1\rvert$	0	0
$\lvert\mathscr{E}_2\rvert$	2	6
$\lvert\mathscr{E}_3\rvert$	2	0

	arch(struct1) $\lvert\mathscr{N}\rvert = 3$
meets ↔ not_meets	
$\lvert\mathscr{E}_1\rvert$	1
$\lvert\mathscr{E}_2\rvert$	0
$\lvert\mathscr{E}_3\rvert$	1
supports ↔ supported_by	
$\lvert\mathscr{E}_1\rvert$	0
$\lvert\mathscr{E}_2\rvert$	0
$\lvert\mathscr{E}_3\rvert$	1

Empirical Study

Types of explanations

- General rule (R): global explanation of the concept a specific instance belongs to
- Example (E): example-based explanation in form of a specific instance belonging to the concept
- Near Miss (N): contrastive (negative) example with high degree of structural similarity to specific instance
- Far Miss (F): contrastive (negative example with low degree of structural similarity to specific instance

Information needs

- Understanding general concept
- Understanding particular example instance for concept
- Understanding what is not in the concept (exclusion)

	(a) Family (R)ule	(E)xample	(N)ear Miss	(F)ar Miss
general	**4.97**	4.52	2.93	2.19
example	4.14	**4.70**	2.49	2.37
exclusion	2.95	2.62	**4.30**	3.67
	(b) Arches (R)ule	(E)xample	(N)ear Miss	(F)ar Miss
general	4.45	**4.70**	2.70	2.36
example	**4.56**	4.27	2.74	2.38
exclusion	3.25	2.73	**3.95**	3.82

Rabold, J., Siebers, M., & Schmid, U. (2021). Generating Contrastive Explanations for Inductive Logic Programming Based on a Near Miss Approach. Machine Learning. Springer