Pruning Hypothesis Spaces Using Learned Domain Theories

Martin Svatoš[1] Gustav Šourek[1] Filip Železný[1]
Steven Schockaert[2] Ondřej Kuželka[2]

[1]Czech Technical University, Prague, Czech Republic

[2]School of CS & Informatics, Cardiff University, Cardiff, UK

September 6, 2017

Motivation

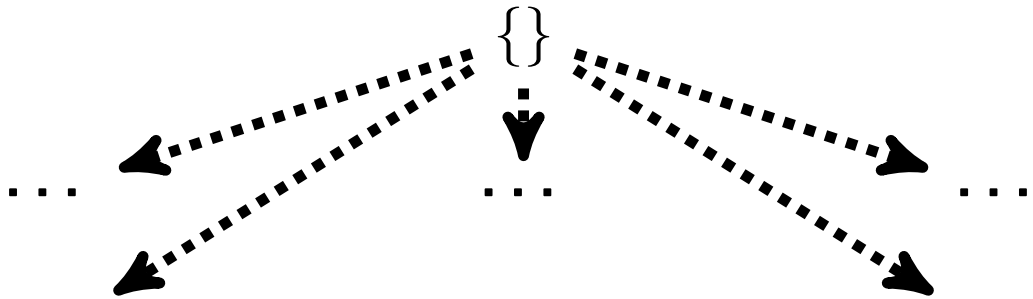
large hypothesis space

Equivalence Testing within Open and Closed List

$C_1 = p_1(A, B) \vee p_2(A, B) \vee p_2(A, C)$
$C_2 = p_1(X, Y) \vee p_2(X, Y) \vee p_2(X, Z)$
$C_3 = p_1(X, Y) \vee p_2(X, Y)$

$C_1 \approx_\theta C_2 \approx_\theta C_3$
$C_1 \approx_{iso} C_2$
C_1 and C_2 are θ-reducible (C_3)

Running Example



$C_4 = oxygen(X) \vee bond(X, Y) \vee fluorine(Y)$

$C_5 = oxygen(X) \vee bond(Y, X) \vee fluorine(Y)$

$\mathcal{B} = \{bond(X, Y) \implies bond(Y, X)\}$

WANTED: $C_4 \approx_{\mathcal{B}} C_5$

Saturation (Definitions)

Let $\mathcal{B}$ be a clausal theory, C a clause and $\mathcal{E}$ a set of examples (without constants or function symbols) .
$e \models \mathcal{B} \quad \forall e \in \mathcal{E}$

If $\mathcal{B} \not\models C$, we define the saturation of C w.r.t. $\mathcal{B}$ to be the maximal clause C' satisfying:

- $vars(C') = vars(C)$, and
- $\mathcal{B} \wedge C'\theta \models C\theta$ for any injective grounding substitution θ.

If $\mathcal{B} \models C$, we define the saturation of C w.r.t. $\mathcal{B}$ to be tautology.

Using Saturations...

Let $\mathcal{B}$ be a clausal theory such that for all examples e from a given dataset it holds that $e \models \mathcal{B}$. Let C be a clause and let C' be its saturation w.r.t. $\mathcal{B}$. Then for any example e from the dataset we have

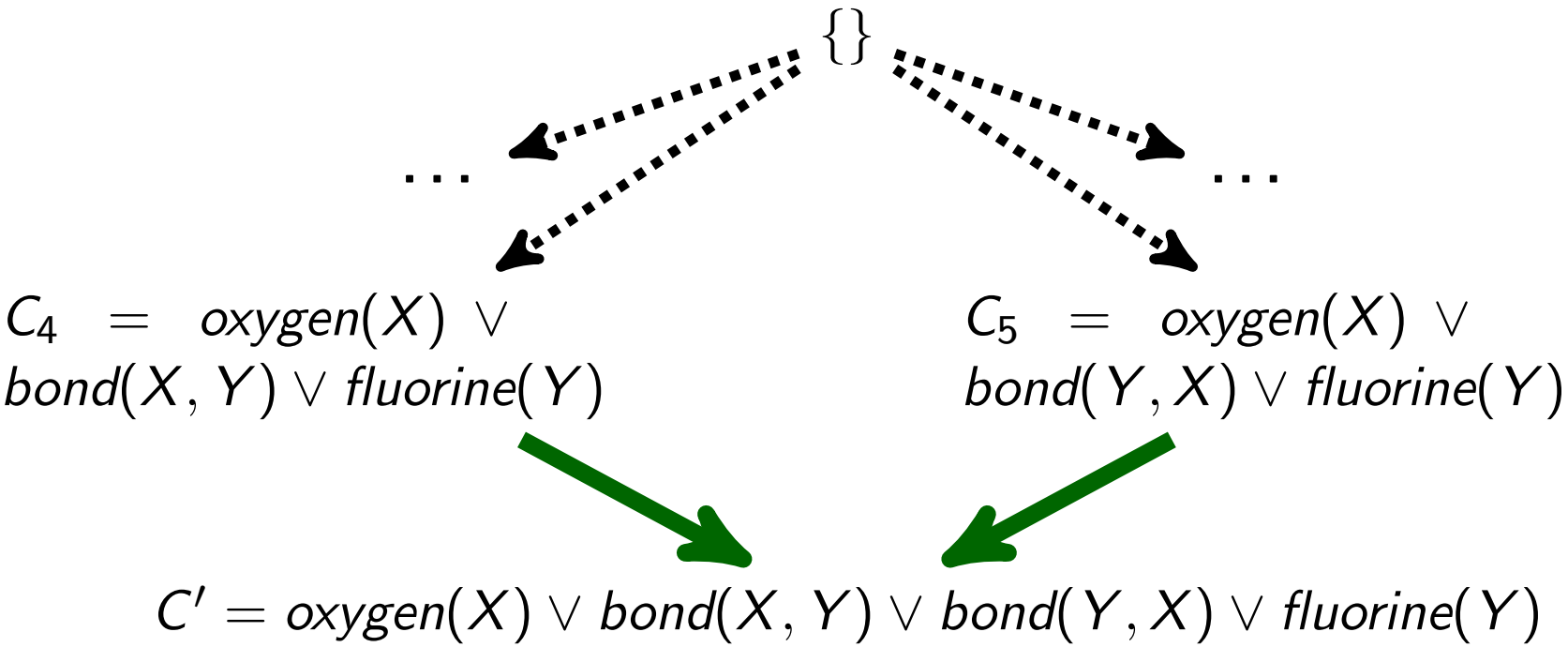
$$(e \models C) \Leftrightarrow (e \models C')$$

Let C_1 and C_2 be clauses and C_1' and C_2' their saturations w.r.t. $\mathcal{B}$, then

$$(C_1 \approx_{\mathcal{B}} C_2) \Leftrightarrow (C_1' \approx_{iso} C_2')$$

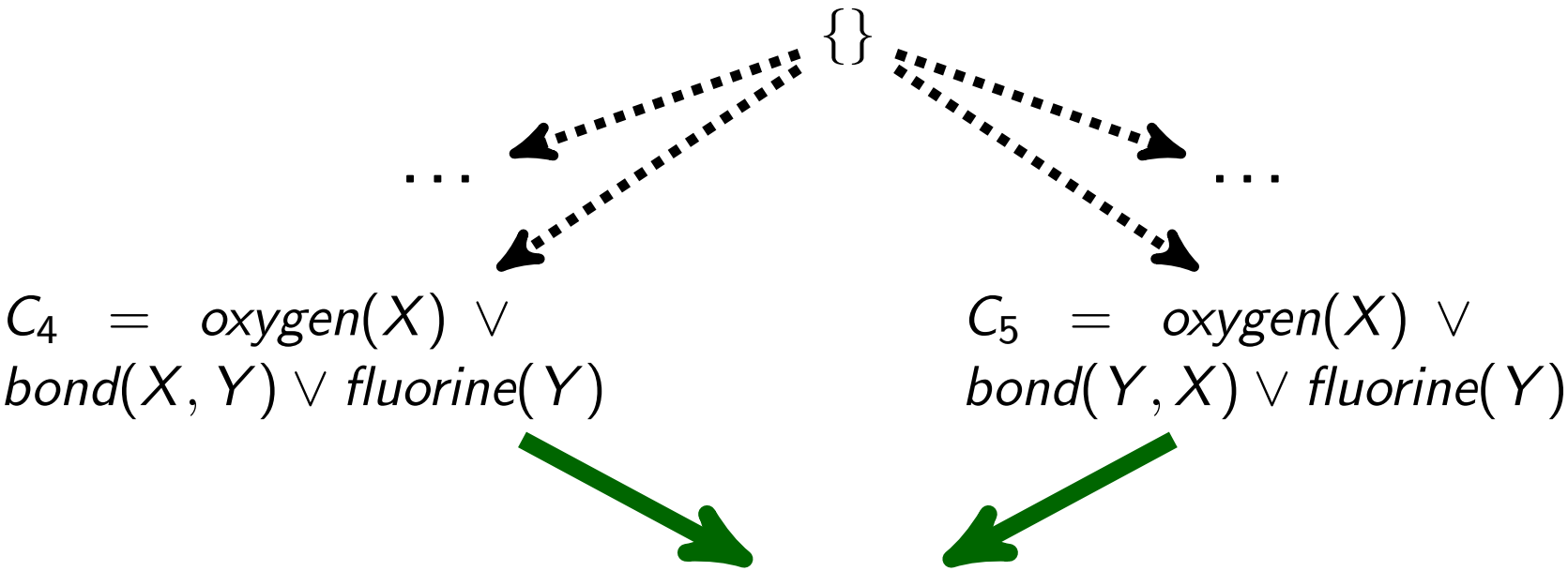
Running Example

$\mathcal{B} = \{bond(X, Y) \implies bond(Y, X)\}$



Running Example

$\mathcal{B} = \{bond(X, Y) \implies bond(Y, X)\}$



$$C_4 \approx_{\mathcal{B}} C_5$$

Using Saturations within Search

Input: domain theory $\mathcal{B}$, candidate clause C, multimaps: open and closed (list) of saturations

$C' \leftarrow saturation(\mathcal{B}, C)$

$hash \leftarrow generalizedWeisfeilerLehman(C')$

$candidates' \leftarrow open.get(hash) \cup closed.get(hash)$

for $candidate' \in candidates'$ **do**

  if $candidate' \approx_{iso} C'$ **then**

    return `// prune clause C`

$open.add(C')$

Generalized Weisfeiler-Lehman labeling procedure used to lower number of isomorphism checking calls.

The pruning preserve completeness of any reasonable ILP search.

Byproduct: Trivial Hypothesis Pruning Using Saturations

$$\mathcal{B} = \{\neg professor(X) \vee \neg student(X)\}$$

$$H = employee(X) \vee \neg professor(X) \vee \neg student(X)$$

$$e \models H \quad \forall e \in \mathcal{E}$$

Learning Domain Theory

$\mathcal{B} \models e \quad \forall e \in \mathcal{E}$

Input: set of samples ($\mathcal{E}$), max length of literals (*maxLiterals*)
Output: domain theory $\mathcal{B}$

$\mathcal{B} \leftarrow \emptyset$
$L_0 \leftarrow \{\{\}\}$
for $i \in \{0 \ldots maxLiterals\}$ **do**
  $L_{i+1} \leftarrow \emptyset$
  for $c \in L_i$ **do**
    for $r \in refinements(candidate, maxLiterals)$ **do**
      if $e \models r \quad \forall e \in \mathcal{E}$ **then** `// refinement r covers all examples`
        if $b \not\preceq_\theta r \quad \forall b \in B$ **then** `// refinement r is not θ-subsumed by any clause from domain theory`
          $\mathcal{B} \leftarrow \mathcal{B} \cup \{r\}$
        else *// discard*
      else $L_{i+1} \leftarrow L_{i+1} \cup \{r\}$
return $\mathcal{B}$

Experiments

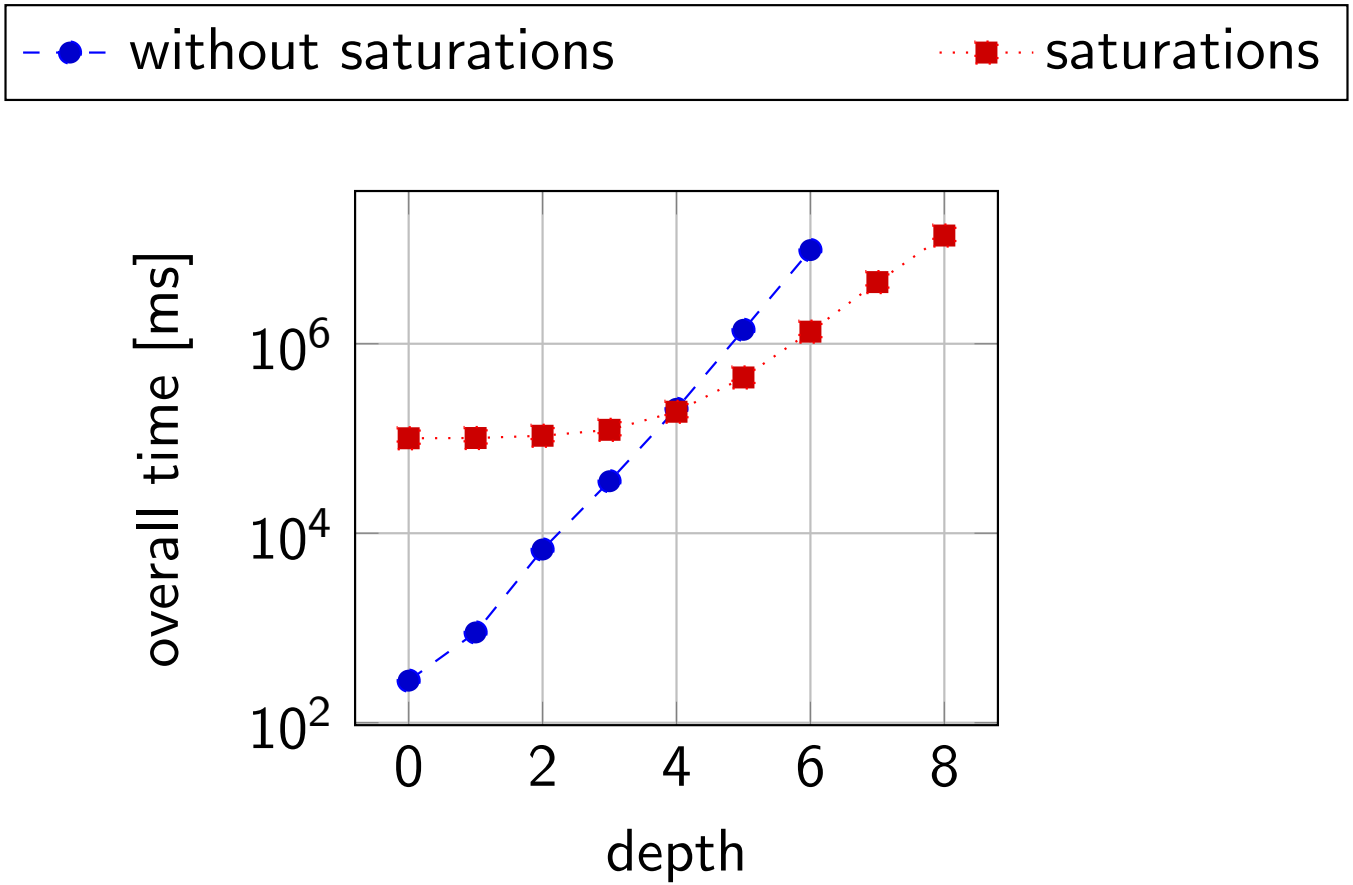


Figure: KM20L2 dataset, 15 hours time limit

Experiments

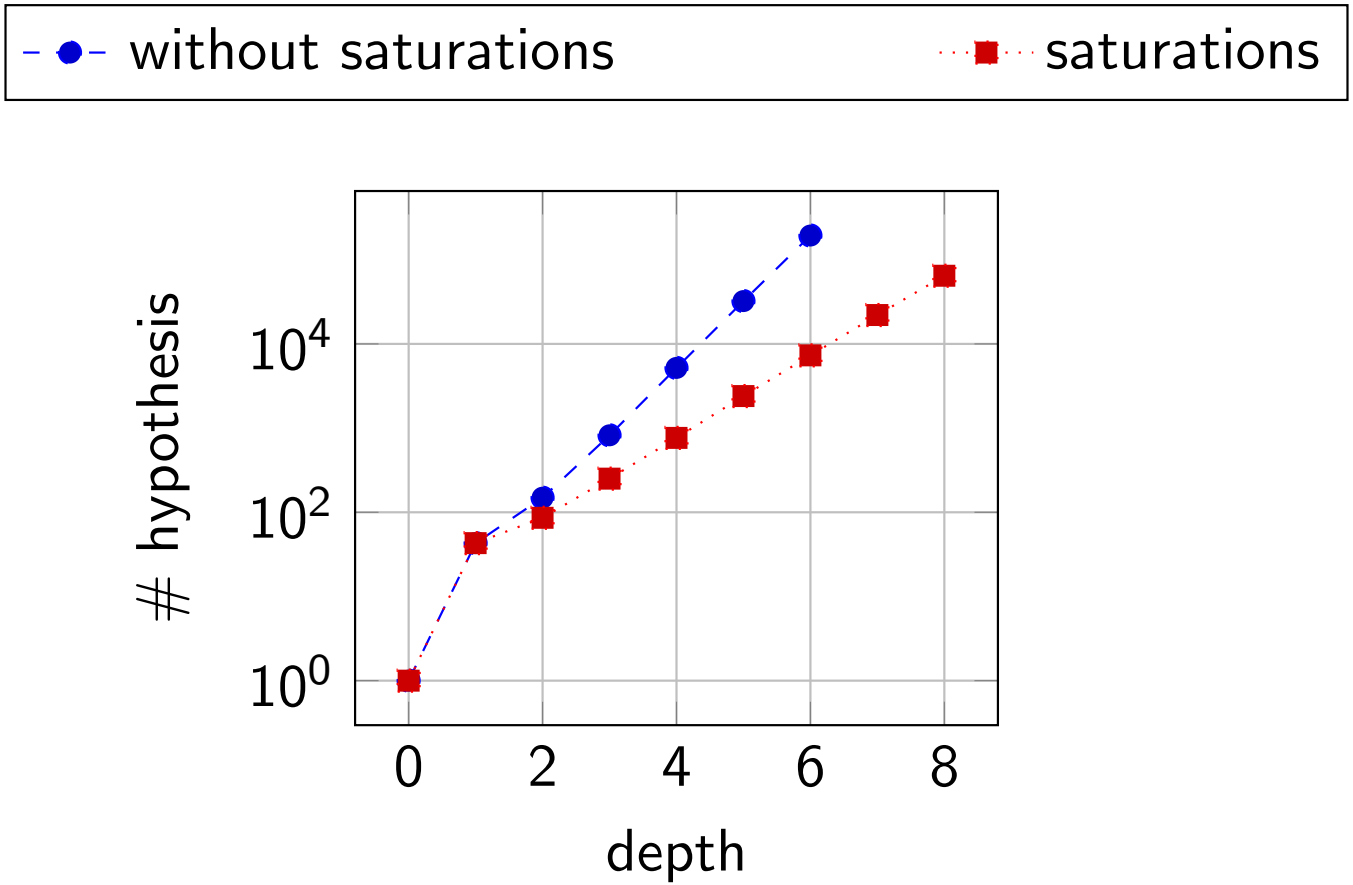


Figure: KM20L2 dataset, 15 hours time limit

Conclusion

- pruning extension of any reasonable ILP search method
- saturations
  - hypotheses equivalence testing relative to domain theory
  - isomorphism checking
- domain theory learner

This work was supported by Czech Science Foundation (17-26999S), Leverhulme Trust (RPG-2014-164) and ERC Starting Grant (637277).