Knowledge Refinement via Rule Selection (AAAI-19)

[♧,◊] Phokion G. Kolaitis, [◊]Lucian Popa, and [◊] Kun Qian
[♧] UC Santa Cruz, [◊] IBM Research – Almaden

Motivation

- Rules (Horn formulas) are ubiquitous in AI

∀x,y,z (PARENT(x,z) ∧ PARENT(y,z) → SIBLING(x,y))
SIBLING(x,y) :- PARENT(x,z), PARENT(y,z)

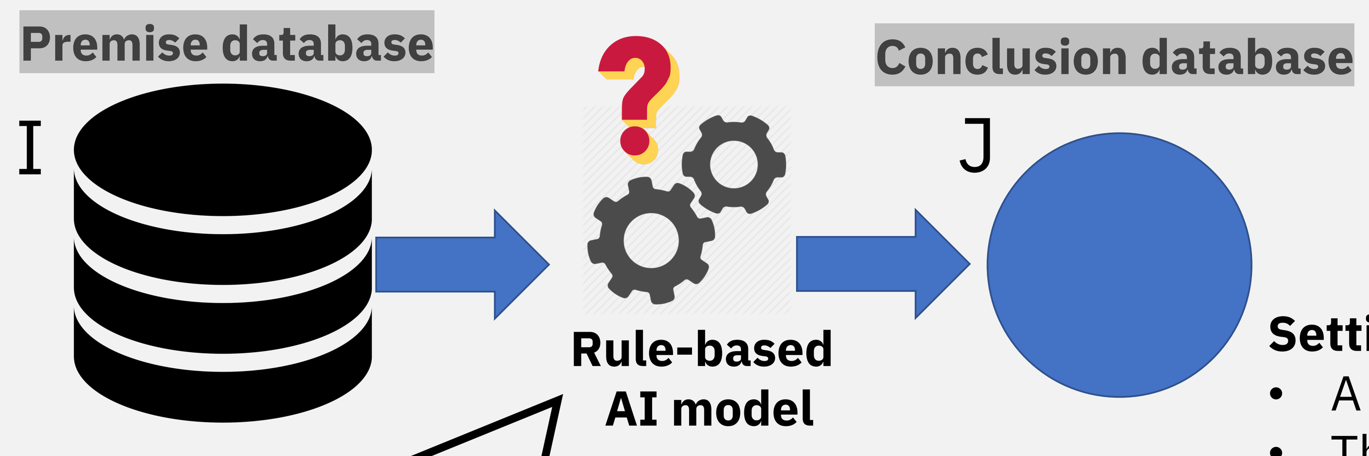
Association rules

TID	Items
1	Cake, Milk
2	Cake, Diaper, Beer
3	Diaper, Beer, Coke
4	Beef, Diaper, Beer
5	Diaper, Bread, Milk, Coke

{Cake} -> {Milk}
{Cake, Diaper} -> {Beer}
{Diaper} -> {Beer}
{Beef, Diaper} -> {Beer}
…… …

Entity resolution rules

Matching publications
DBLP(id, **title, authors, venue,** year)
∧ ACM(id', **title, authors, venue,** year') → Same(id,id')



A large set C of possible rules for a given task exists

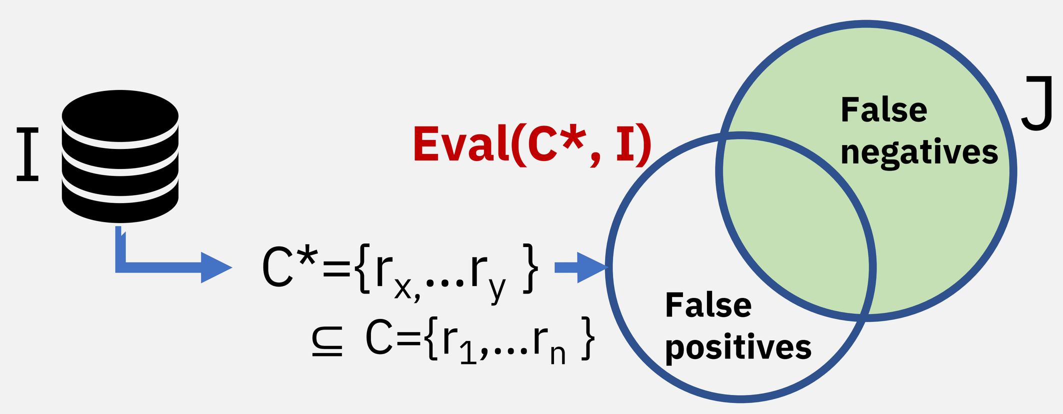
R7: dblp(id1, title, …) -> same(id1, id2)
R8: dblp(id1, title, …) -> same(id1, id2)
R9: dblp(id1, title, …) -> same(id1, id2)
R10: dblp(id1, title, …) -> same(id1, id2)
R11: dblp(id1, title, …) -> same(id1, id2)
R12: dblp(id1, title, …) -> same(id1, id2)
… …. …. … …

R1: dblp(id1, title, …) -> same(id1, id2)
R2: dblp(id1, title, …) -> same(id1, id2)
R3: dblp(id1, title, …) -> same(id1, id2)
R4: dblp(id1, title, …) -> same(id1, id2)
R5: dblp(id1, title, …) -> same(id1, id2)
R6: dblp(id1, title, …) -> same(id1, id2)
… …. …. … …

Setting:
- A large set C of possible rules is available.
- The actual semantics of the task is given via data: a premise database I together with a conclusion database J that represents the expected output of the task on input I.

Question: Can we capture the semantics via a small subset of C?

Min Rule Selection Problem



- **Total error** = false positives + false negatives

Definition. Min Rule-Select Problem

Given a pair (I, J) of a premise and a conclusion database, and a set C of rules, find a subset C* of C such that the total error of Eval(C*, I) w.r.t. J is minimized.

Theorem 1

Min Rule-Select is an NP-hard optimization problem.

Hence, unless NP=P, there is no polynomial time algorithm that solves the problem exactly.

? Are there PTIME approximation algorithms with formal guarantees?

Theorem 2

Min Rule-Select is approximable within a factor of $2\sqrt{|C| + |J| \cdot \log |J|}$, where $|C|$ is the number of input rules and $|J|$ is the size of the conclusion instance.

We show that Min Rule-Select is "equivalent" to the Positive-Negative Partial Set-Cover Problem [Miettinen, 2008]

Bi-objective and Bi-level Optimization

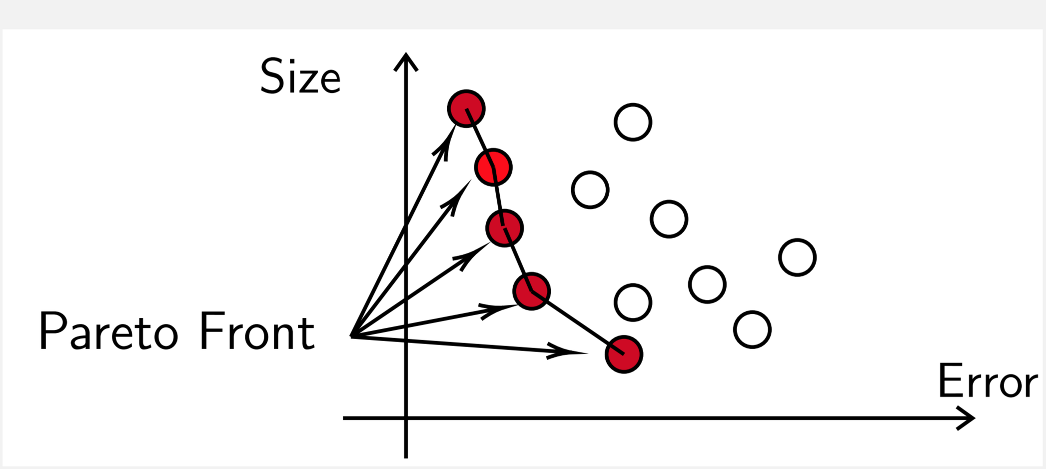
? What if we want to optimize both the *error* and the *size* of the set of rules? (Note that *error* and *size* are incomparable quantities.)

Definition. Pareto Optimal Solution

Given a set C of rules and a pair (I,J) of a premise and a conclusion database, a subset C* of C is a ***Pareto optimal*** solution if there is no subset C' of C such that size(C') < size(C*) and error(C')<error(C*).
The ***Pareto front*** is the set of all pairs (s*, e*) of integers such that there is a Pareto optimal solution C* with size(C*)=s* and error(C*)=e*.

Definition. Bi-level Optimal Solution

Given a set C of rules and a pair (I,J) of a premise and a conclusion database, a subset C* of C is a ***Bi-level optimal*** solution if it has both minimum error and minimum size among all minimum-error solutions.



Theorem 3

The following problems are coNP-complete:
Given a pair (I,J) of a premise and a conclusion database, a set C of rules, and a subset C* of C:
- *is C* a Pareto optimal solution?*
- *is C* a bi-level optimal solution?*

Summary of Results

		FP	FP+FN
Rule-Select		NP-complete	NP-complete
Exact Rule-Select		DP-complete	DP-complete
Min Rule-Select	approx. upper	$2\sqrt{\lvert\mathcal{C}\rvert \log \lvert J\rvert}$	$2\sqrt{(\lvert\mathcal{C}\rvert + \lvert J\rvert) \log \lvert J\rvert}$
	approx. lower	$2^{\log^{1-\epsilon}(\lvert\mathcal{C}\rvert)}, \forall\epsilon > 0$	$2^{\log^{1-\epsilon}(\lvert J\rvert)}, \forall\epsilon > 0$
Pareto Optimal Solution		coNP-complete	coNP-complete
Pareto Front Membership		DP-complete	DP-complete
Bi-level Optimal Solution		coNP-complete	coNP-complete
Bi-level Optimal Value		DP-complete	DP-complete

Directions for future work

- Experimental evaluation of the approximation algorithms for Min Rule-Select.
- Heuristics or approximation algorithms for constructing the Pareto front of rule selection.