Knowledge Refinement via Rule Selection

Phokion G. Kolaitis[1,2], Lucian Popa[2], Kun Qian[2]

[1]UC Santa Cruz, [2]IBM Research – Almaden

AAAI-19

UNIVERSITY OF CALIFORNIA SANTA CRUZ

IBM Research

Introduction

- Rules are ubiquitous in AI and computer science
- Rules as typically specified by Horn formulas
  $\forall x, y, z(\mathrm{PARENT}(\mathrm{x}, \mathrm{z}) \wedge \mathrm{PARENT}(\mathrm{y}, \mathrm{z}) \rightarrow \mathrm{SIBLING}(\mathrm{x}, \mathrm{y}))$
- Rules are extensively used in different areas, including:
  - Logic Programming
    $\mathrm{SIBLING}(\mathrm{x}, \mathrm{y}) \;\; :- \; \mathrm{PARENT}(\mathrm{x}, \mathrm{z}), \mathrm{PARENT}(\mathrm{y}, \mathrm{z})$
  - Data Mining (Mining Association Rules)
  - Data Exchange and Data Integration
  - Entity Resolution.
- Central Problem: From a large number of possible rules, select a subset that best captures the task at hand.

Concrete Scenario - Refining Entity Resolution Rules

- Discover a rule-based model that identifies the same paper across datasets

  DBLP: (id, title, authors, venue, year), ACM: (id, title, authors, venue, year)

- There is a large number of possible entity resolution rules:

r1:	DBLP(p, **t, a, v**, y1)	$\wedge$ ACM(q, **t, a, v**, y2)	$\rightarrow$ Same(p, q)
r2:	DBLP(p, **t**, a1, **v, y**)	$\wedge$ ACM(q, **t**, a2, **v , y**)	$\rightarrow$ Same(p, q)
r3:	DBLP(p, t1, a1, **v, y**)	$\wedge$ ACM(q, t2, a2, **v, y**)	$\rightarrow$ Same(p, q)
...		...	...

- Rules behave differently.
  - `r1` has high precision but low coverage.
  - `r3` has high coverage but low precision.
- Goal: Find an optimal subset of the given rules that achieves high precision and high coverage.

Concrete Scenario - Refining Entity Resolution Rules

- Given a set of available rules and ground truth data:

r1: DBLP(p, **t, a, v**, y1)	∧ ACM(q, **t, a, v**, y2)	→ Same(p, q)
r2: DBLP(p, **t**, a1, **v, y**)	∧ ACM(q, **t**, a2, **v , y**)	→ Same(p, q)
r3: DBLP(p, t1, a1, **v**, y1)	∧ ACM(q, t2, a2, **v**, y2)	→ Same(p, q)
...	...	...

DBLP records	ACM records
(d_1, "rule selection", "Jone Doe", AAAI, 2019)	(a, "rule selection", "J. Doe", AAAI, 2019)
(d_2, "Invited Talk", "Qiang Yang", AAAI, 2019)	(b, "Invited Talk", "Yu Zheng", AAAI, 2019)
(d_3, "Keynote", "Jane Doe", SIGMOD, 2018)	(c, "Keynote", "Jane Doe", SIGMOD, null)

Same
(d_1,a)
(d_3,c)

- Which model minimizes the error over the ground truth?
  - {r1}: 0 false positives, 1 false negative.
  - {r2}: 1 false positives, 1 false negative.
  - {r3}: 2 false positive, 0 false negatives.
  - Both {r1} and {r1, r2} achieve minimum error.

Basic Concepts: Rules, Data Examples, Rule Evaluation

- We consider rules of the form $\forall \mathbf{x}\big(\psi(\mathbf{x}) \rightarrow P(\mathbf{x})\big)$, where
  - the premise $\psi(\mathbf{x})$ is a conjunction of atoms over a schema **S**;
  - the conclusion $P(\mathbf{x})$ is a atom over a schema **T** disjoint from **S**.
- A data example is a pair (I, J), where I is a premise instance I over **S** and a J is a conclusion instance over **T**.
- If $\mathcal{C}$ is a set of rules, then $\text{Eval}(\mathcal{C}, I) = \bigcup_{\mathtt{r} \in \mathcal{C}} \text{Eval}(\mathtt{r}, I)$ (i.e., evaluate the premise of each rule on I and populate the conclusion).

Premise Instance I

DBLP records	ACM records
(d_1,"rule selection","Jone Doe",AAAI,2019)	(a,"rule selection","J. Doe",AAAI,2019)
(d_2,"Invited Talk","Qiang Yang",AAAI,2019)	(b,"Invited Talk","Yu Zheng",AAAI,2019)
(d_3,"Keynote","Jane Doe",SIGMOD,2018)	(c,"Keynote","Jane Doe",SIGMOD,null)

Conclusion Instance J

Same
(d_1,a)
(d_3,c)

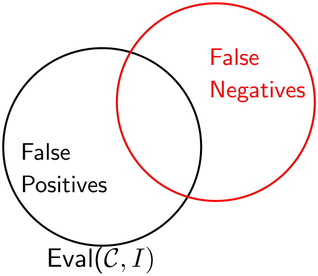
$\text{Eval}(\{\mathtt{r}_1\}, I) = \{(d_3, c)\}$, $\text{Eval}(\{\mathtt{r}_1, \mathtt{r}_2\}, I) = \{(d_1, a), (d_2, b), (d_3, c)\}$

False Positive and False Negative Errors

Definition (Errors)

Given a set $\mathcal{C}$ of rules and a data example (I, J), we consider the following two types of errors of $\mathcal{C}$ w.r.t. (I, J):

- False positive errors: $\mathsf{Eval}(\mathcal{C}, I) \setminus J$;
- False negative errors: $J \setminus \mathsf{Eval}(\mathcal{C}, I)$.



The Min Rule-Selection Problem

Definition (MIN RULE-SELECT$_{\mathsf{FP+FN}}$)

Input: A set $\mathcal{C}$ of rules and a data example (I, J).
Goal: Find a subset $\mathcal{C}^* \subseteq \mathcal{C}$ such that the sum of the number of the false positive errors and the number of false negative errors of $\mathcal{C}^*$ with respect to (I, J) is minimized.

Definition (MIN RULE-SELECT$_{\mathsf{FP}}$)

Input: A set $\mathcal{C}$ of rules and a data example (I, J) with $J \subseteq \mathsf{Eval}(\mathcal{C}, I)$.
Goal: Find a subset $\mathcal{C}^* \subseteq \mathcal{C}$ such that the number of false negative errors is zero and the number of false positive errors of $\mathcal{C}^*$ with respect to (I, J) is minimized.

Note: MIN RULE-SELECT$_{\mathsf{FN}}$ has a trivial solution (the entire set of rules).

Complexity of MIN RULE-SELECT Problems

Theorem

Both MIN RULE-SELECT$_{\mathsf{FP+FN}}$ and MIN RULE-SELECT$_{\mathsf{FP}}$ are NP-hard optimization problems.

Hint of Proof: Reduction from SET COVER.

Question:

Are there polynomial-time approximation algorithms for MIN RULE-SELECT$_{\mathsf{FP+FN}}$ and MIN RULE-SELECT$_{\mathsf{FP}}$?

Approximation Properties of $\textsc{Min Rule-Select}_{\mathsf{FP}}$

Theorem

- $\textsc{Min Rule-Select}_{\mathsf{FP}}$ is approximable within a factor of $2\sqrt{|\mathcal{C}| \log |J|}$, where $|\mathcal{C}|$ is the number of input rules and $|J|$ is the size of the conclusion instance J.
- Unless P=NP, for every $\epsilon > 0$, there is no polynomial time algorithm that approximates $\textsc{Min Rule-Select}_{\mathsf{FP}}$ within a factor of $2^{\log^{1-\epsilon}(|\mathcal{C}|)}$, where $|\mathcal{C}|$ is as above.

Hint of Proof: Give approximation-preserving reductions between $\textsc{Min Rule-Select}_{\mathsf{FP}}$ and the $\textsc{Red-Blue Set Cover}$ problem, which was studied by [Peleg, 2007].

Approximation Properties of $\textsc{Min Rule-Select}_{\mathsf{FP+FN}}$

Theorem

- $\textsc{Min Rule-Select}_{\mathsf{FP+FN}}$ is approximable within a factor of $2\sqrt{(|\mathcal{C}| + |J|)\log|J|}$, where $|\mathcal{C}|$ is the number of input rules and $|J|$ is the size of the conclusion instance J.
- Unless P=NP, for every $\epsilon > 0$, there is no polynomial time algorithm that approximates $\textsc{Min Rule-Select}_{\mathsf{FP+PN}}$ within a factor of $2^{\log^{1-\epsilon}(|J|)}$, where $|J|$ is as above.

Hint of Proof: Give approximation-preserving reductions between $\textsc{Min Rule-Select}_{\mathsf{FP+FN}}$ and the $\textsc{Positive-Negative Partial Set Cover}$ problem, which was studied by [Miettinen, 2008].

Bi-objective Optimization Problems

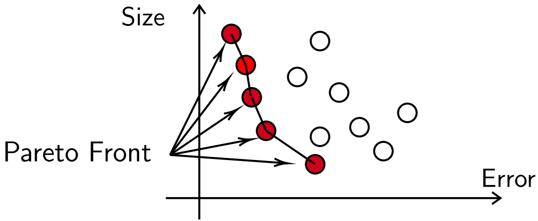
Question: What if we want to optimize both the errors and the size of rules simultaneously?

- Note that error and size are qualitatively incomparable quantities.

Answer: Consider bi-objective optimization problems.

Definition (Pareto Optimality)

Given a set $\mathcal{C}$ of rules and a data example (I, J), a subset $\mathcal{C}^* \subseteq \mathcal{C}$ is a Pareto optimal solution if there is no $\mathcal{C}' \subseteq \mathcal{C}$ with size$(\mathcal{C}')$ $<$ size$(\mathcal{C}^*)$ and error$(\mathcal{C}')$ $<$ error$(\mathcal{C}^*)$.



Complexity Results for Bi-objective Optimization

Theorem

- The following problem is coNP-complete: given an instance $K = \{\mathcal{C}, (I, J)\}$ of $\textsc{Min Rule-Select}_{\mathsf{FP}}$ and a subset $\mathcal{C}^* \subseteq \mathcal{C}$, is $\mathcal{C}^*$ a Pareto optimal solution?
- The following problem is DP-complete: given an instance $K = \{\mathcal{C}, (I, J)\}$ of $\textsc{Min Rule-Select}_{\mathsf{FP}}$ and pair (s, e) of integers, is (s, e) on the Pareto front of K?

Similar results hold for $\textsc{Min Rule-Select}_{\mathsf{FP+FN}}$.

Corollary

Unless P $=$ NP, there is no polynomial-time algorithm for constructing the Pareto front.

Concluding Remarks

This work establishes foundations and complexity results for the rule-selection problem, which is fundamental in many areas of AI.

A summary of our results (more details can be found in the paper).

		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

- Approximating the Pareto front of the rule selection problem.
- Experimental evaluation.

Thank you

References I

Miettinen, P. (2008).
On the positive-negative partial set cover problem.
Information Processing Letters, 108(4):219 – 221.

Peleg, D. (2007).
Approximation algorithms for the label-covermax and red-blue set cover problems.
Journal of Discrete Algorithms, 5(1):55 – 64.