PU-learning disjunctive concepts in ILP

Hendrik Blockeel
Dept. of Computer Science, KU Leuven
Motivation

• “Relational grounded language learning” (Becerra-Bonache et al., IDA 2015, ECAI 2016)

• Learn meaning of phrases (“n-grams”) from occurrences in some context

• Mentioned $\Rightarrow$ occurs, but occurs $\nRightarrow$ mentioned
Motivation

• Earlier work: “Meaning” of an n-gram = least general generalization under subsumption (lgg, Plotkin 1970) of all contexts where it occurred. Intuitively: “maximal common pattern”

  • E.g.: “mike” -> object(X), human(X,boy), pose(X, _), expression(X, _)

  • E.g. “mike is wearing a hat” ->
    object(A), human(A, c_boy), pose(A, _), expression(A, _), object(D), clothing(D, c_hat), style(D, _), act(A, c_wear, D)

• This does not allow for disjunctive concepts

• Those are more common than expected! E.g.: “dog”
PU-learning

- “Normal” supervised concept learning: learn from positive and negative data
  - Ideal rule set covers all positives and no negatives
- PU-learning: Learn from positive and unlabeled data
  - Ideal rule set covers all positives and some unlabeled (we don’t know which, nor how many)
Supervised learning
PU learning
PU-learning

• Recall: **Mentioned** ⇒ **occurs**, but **occurs** ⇒ **mentioned**

• Viewing mentions as the labels, this makes our setting **PU-learning**

  • “Cat” in sentence labels the context as positive for cat

  • No “cat” in sentence does *not* label the context as negative, just unlabeled for cat
Elkan & Noto's result

• Elkan & Noto (2008) observed:
  
  • Let $L = \text{``x is labeled as positive''}$, $+/\overline{\text{}} = \text{``x is positive/negative''}$
  
  • Assume positives are labeled completely at random: $P(L|+, x) = P(L|+) = k$
  
  • Then $P(L|x) = P(L|+) P(+|x) + P(L|-) P(-|x) = cP(+|x) + 0P(-|x) = kP(+|x)$
  
  • From a probabilistic classifier that predicts $L$, we can derive a probabilistic classifier that predicts $+$, if we know $k$
  
  • The former can be learned in a supervised manner
  
  • PU-learning reduces to supervised learning under the mentioned assumptions
  
  • Ways to estimate $k$ have been proposed (e.g. Bekker & Davis, ILP 2017)
Weakening Elkan & Noto’s “constant c” assumption

- Learning disjunctive concepts with constant $c = P(\text{label}|\text{pos})$: see Bekker & Davis, ILP 2017 (previous talk)

- But “constant c” is not realistic in our setting

- e.g. $P(\text{“dog” | dog occurs}) \neq P(\text{“dog” | hot dog occurs})$

  $\frac{538}{1706}=0.315 \quad \frac{134}{769}=0.174$

- More “noteworthy” things are more likely to be mentioned
PU-learning disjunctive concepts

• Consider the following setting:

• The concept is disjunctive: \( d_1 \lor d_2 \lor \ldots \lor d_k \)

• Assumption: Within each disjunct \( d_i \), each \( x \) has the same probability \( c_i \) of being labeled, but it is possible that \( c_i \neq c_j \) for \( i \neq j \)

• How to PU-learn in this setting?

• Estimate \( c_i \) for a given disjunct \( d_i \)? Then we first need to know \( d_i \)! Catch-22. Need to learn \( d_i \) and \( c_i \) simultaneously.
Our proposal

- Learn one rule (= disjunct) at a time, using a bottom-up rule learner (BURL)
  - In our case: Golem-like learner (Muggleton & Feng 1992), generalizing via lgg

- Normally, BURL starts with a specific rule (covering 1 example), and generalizes until it can’t generalize anymore without including negatives

- In PU-learning, the stopping criterion is less obvious: generalization should include unlabeled - but not too many. What’s too many?
Assume 2 disjuncts $d_1$, $d_2$ to be learned.

$P(L|d_1) = c_1$

$P(L|d_2) = c_2$
Start with a seed example, generalize!
As long as rule $\subseteq$ disjunct, $P(L|\text{rule})$ is constant!
P(L|\text{rule}) \text{ drops when too general} \\
P(L|d_1)=c_1 \\
P(L|d_2)=c_2 \\
P(L|\text{rule})<c_1
With more disjuncts, $P(L|\text{rule})$ may change in any direction.

$$P(+|d_1) = c_1$$

$$P(+|d_2) = c_2$$

$$P(L|\text{rule}) = c_1w_1 + c_2w_2$$
Assumption

• Our method implicitly assumes that $P(L|\text{rule})$ starts going down when $\text{rule} \not\subseteq c_1$

• This assumption holds when $c_1 > c_2$, or when disjuncts are “small” and “faraway”
While rule $\subseteq d_1$: $P(L|\text{rule}) = c_1$

As soon as rule $\not\subseteq d_1$: $P(L|\text{rule}) < c_1$
... But we have estimates for P, not P itself!
Estimates are closer to P as coverage increases
We can construct confidence intervals. To find the rule with largest coverage \( \subseteq d_1 \), choose the point with maximal lower bound. This heuristic makes sense because

- Very small coverage \( \Rightarrow \) wide interval \( \Rightarrow \) low LB
- Rule \( \not\subseteq d_1 \Rightarrow \) low expectation \( \Rightarrow \) low LB
Algorithm: PULOR  
(PU-Learn One Rule)

• Choose a random positive $e$

• $d_0 = e$, $i=0$

• While examples left:
  • Choose $\leq s$ random examples $e_j$, not covered by $d_i$
  • $d_{i+1} = \text{argmax}_j \text{ quality}(\text{lgg}(d_i, e_j))$
  • $i++$

• Return $\text{argmax}_i \text{ quality}(d_i)$

With $\text{quality}(d) = \text{LB}(P(\text{pos}|d))$
Algorithm: PULSE

• Repeat

• choose a random uncovered positive example, e

• $R = \text{PULOR}(e)$

• Add $R$ to RuleSet

• Mark positive examples covered by $R$ as covered

• Until no good rule $R$ can be found

• Remove redundant rules (subsumed by other rules)
Experiments on a dataset with 10k (sentence, context) pairs (Becerra-Bonache et al., ECAI 2016, derived from Zitnick et al.’s (2013) dataset): some representative results

<table>
<thead>
<tr>
<th>Cat</th>
<th>[object(A), animal(A, c_cat), size(A, c_small)]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dog</td>
<td>[object(A), animal(A, c_dog), color(A, c_brown), size(A, c_small)]</td>
</tr>
<tr>
<td></td>
<td>[object(A), color(A, B), object(C), food(C, c_hot_dog)]</td>
</tr>
<tr>
<td>Table</td>
<td>[object(A), large(A, c_table), color(A, c_yellow)]</td>
</tr>
<tr>
<td>Sitting</td>
<td>[object(A)]</td>
</tr>
<tr>
<td>The dog</td>
<td>[object(A), animal(A, c_dog), color(A, c_brown), size(A, c_small)]</td>
</tr>
<tr>
<td>Hot dog</td>
<td>[object(A), food(A, c_hot_dog)]</td>
</tr>
<tr>
<td>Hot air balloon</td>
<td>[object(A), sky(A, c_hot_air_balloon), size(A, c_big), act(A, c_fly)]</td>
</tr>
<tr>
<td>Mike</td>
<td>[object(A), human(A, c_boy), pose(A, B), expression(A, C)]</td>
</tr>
<tr>
<td>Angry</td>
<td>[object(A), color(A, B), object(C), human(C, c_boy), pose(C, D), expression(C, c_angry)]</td>
</tr>
<tr>
<td></td>
<td>[object(A), color(A, B), object(C), human(C, c_girl), pose(C, D), expression(C, c_angry)]</td>
</tr>
</tbody>
</table>
Quality curves

“Hot dog”
Quality curves

Quality

```
[object(A),large(A,B),object(C),object(D),sky(C,E),color(D,F),size(C,c_big),object(G),animal(G,c_cat),size(G,c_small)]
```

```
[object(A),animal(A,c_cat),size(A,c_small)]
```

# generalization steps

“Cat”
Quality curves

“Dog”
Conclusions

• Main contribution: PULSE: A new method for PU-learning rule sets in ILP, does not assume constant $P(L|+)$

• Applied to relational grounded language learning:
  • Finds similar conjunctive concepts as before, but can also identify disjunctive concepts
  • Interpretable results obtained
  • “Quality curves” confirm soundness of method’s principle
  • (Bonus: ad-hoc noise handling method, used in earlier work, now no longer needed)