Toward Statistical Predicate Invention

Stanley Kok & Pedro Domingos

Dept. of Computer Science and Eng.

University of Washington

Predicate Invention

- **Inductive Logic Programming (ILP)** approach
- Form predicates to represent
  - *a.* Commonalities (interconstruction) [Wogulis & Langley, 1989]
  - *b.* Differences (intraconstruction) [Muggleton & Buntine, 1988]
    - § *a.* and *b.* prone to over-generating predicates
  - *c.* Exceptions to rules [Srinivasan et. al., 1992]
- Form predicates from 2nd-order templates [Silverstein & Pazzani, 1991]
- **Limited ability to handle noisy data**

Latent/Hidden Variable Discovery

- **Statistical learning** approach
- Form hidden variables from
  - *a.* Structural patterns in Bayesian networks [Elidan et. al., 2001]
  - *b.* Observed variables grouped by mutual information [Elidan & Friedman, 2005]
- EM algorithm iteratively
  - *a.* Creates hidden variables
  - *b.* Hypothesizes hidden variables values
  - *c.* Learns parameters of resulting Bayesian network
- **Assumes data independent and identically distributed**

Statistical Predicate Invention (SPI)

What is SPI?

- Discovery of new concepts, relations, properties
- Combines ILP and statistical learning approaches
- Invented predicates →discover more predicates
- More powerful than learning from fix set of primitives

Benefits

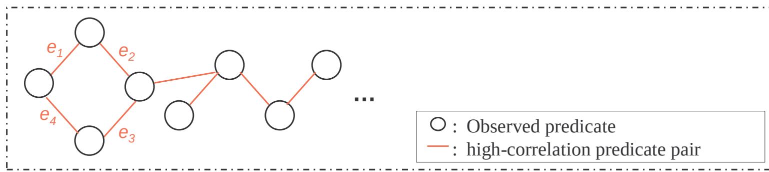
- More compact and comprehensible model
- Reduce # parameters from exponential to linear
- Reduce risk of overfitting
- Less memory
- Potentially faster inference
- Invented predicate used to learn new formulas
  => larger steps through search space
- Represent unobserved aspects => better accuracy

Applications	Invented Predicates
Activity Recognition	•High-level activity (e.g., cooking, taking medication) •Daily routines from high-level activities
Robotics	Corridors, doorways, etc.
Perception (speech/handwritg recognition)	•Parts of objects •Objects as related set of parts
Molecular biology	• Gene modules • Metabolic pathways • Cell substructures
Security	•Steps of criminals' plan •Relations among steps •Criminal's roles
Many more…	…

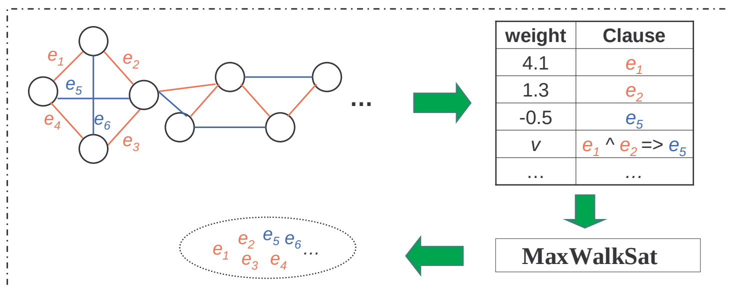
Proposed Approach

1a. Compute correlations of all pairs of predicates (all variabilizations)

1b. Discard low correlation pairs

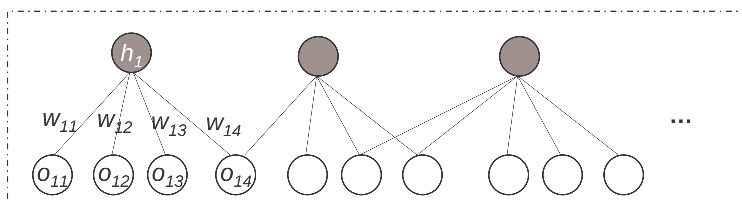


2. Find clusters of predicates that are highly correlated
   - Express as weighted satisfiability problem
   - Each pair of predicates is an atom and unit clause:
     Edge(P1,P2) with weight = log(|correlation(P1,P2)|) - *thresh*
   - Apply "soft" transitive closure:
     Edge(P1,P2) ^ Edge(P2,P3) => Edge(P1,P3) with weight *v*
     Higher *v* => Larger clusters of predicates
   - Use MaxWalkSat [Kautz et. al. 1997] to solve sat. problem & select edges



3a. Invent a predicate for each clique of predicates
- Arguments are (a subset of) the observed predicates' arguments

3b. Model correlation among predicates in clique
- Associate a weight w_{ij} between invented predicate h_i and each of its observed predicate o_{ij}



4a. Define a potential f_{ijk} between the k^{th} grounding of invented predicate h_i and each of its observed predicate o_{ij}

4b. When the invented predicates are independent given the observables, we can sum them out and avoid using EM

5a. Init weights w_{ij} to the average (log) correlation between o_{ij} and other observed predicates of h_i

5b. Find locally optimal weights using gradient ascent

6. Iterate by treating the hidden predicates as observed predicates, and setting them to their MAP values

$$f_{ijk} = \begin{cases} 1 & \text{if } o_{ijk} = 1 \text{ \& } h_{ik} = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$Q_i(x, h_{ik}) = \exp(\sum_j w_{ij} f_{ijk})$$

$$P(x) = \frac{1}{Z} \prod_{i \in IP} \prod_{k \in G_i} (Q_i(x, h_{ik} = 1) + 1)$$

where Z is a partition function,
IP is a set of invented predicates,
G_i is a set of all groundings of invented predicate h_i, and
h_{ik} is the value of the k^{th} grounding of h_i.