Learning Rules With Categorical Attributes from Linked Data Sources

Andre de Oliveira Melo

Saarland University

andresony@gmail.com

January 15, 2013

Overview

Semantic Web

> *"provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries"*

Built on W3C's:

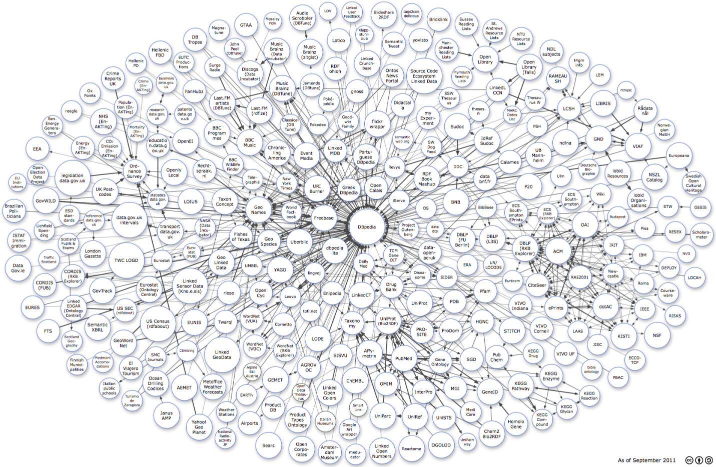
- RDF
- OWL
- SKOS
- SPARQL

Linked Data

"a term used to describe a recommended best practises for exposing, sharing, and connecting pieces of data, information and knowledge on the Semantic Web using URIs and RDF"

"collection of interrelated datasets on the Web"

Linked Data



Motivation

Learn inference rules from data:

$$\underbrace{livesIn(x, y)}_{head}\text{:-}\underbrace{isMarriedTo(x, z), livesIn(z, y)}_{body}$$

Support and confidence thresholds

- Support: $supp(head\text{:-}body) = supp(head \cup body)$
- Confidence: $conf(head\text{:-}body) = \dfrac{supp(head \cup body)}{supp(body)}$

Motivation

Introducing constants can be relevant, e.g.:

$$\textit{speaks(x,y) :- livesIn(x,z)}$$
$$\textit{speaks(x,Portuguese) :- livesIn(x,Brazil)}$$

What about numerical attributes?

$$\textit{hasChild(x,y) :- hasAge(x,a)}\ \text{[base-rule]}$$

- **Support**: number of supporting examples
  $supp(head\text{:-}body) = supp(head \cup body)$
- **Confidence**: $conf(head\text{:-}body) = \dfrac{supp(head \cup body)}{supp(body)}$

Motivation

We are more interested in base-rules that:

- Satisfy support threshold
- Do not satisfy confidence threshold
- Potentially has a refined-rule with an interval that satisfies both thresholds

  i.e., has non-uniform confidence distribution

  i.e., has divergent positive examples and body support distributions

Motivation

Problem?

- Search space grows dramatically
- Usually unfeasible to perform exhaustive search
- Querying support and confidence distributions is very expensive

ILP

Inductive Logic Programming: Finds a hypothesis H that covers all positive, and no negative examples

$$positiveExamples + negativeExamples + backgroundKnowledge \rightarrow hypothesis \quad (1)$$

Training Examples	**Background Knowledge**
daughter(mary,ann) +	parent(ann,mary)
daughter(eve,tom) +	parent(ann, tom)
daughter(tom,ann) -	parent(tom,eve)
daughter(eve,ann) -	parent(tom,ian)
	female(ann)
	female(mary)
	female(eve)

Important concepts:

- Literal: predicate symbol with bracketed n-tuple, e.g:
  $L = livesIn(x, y)$
- Clause: a disjunction of literals (negated or not), e.g:
  $c = (L_1 \vee L_2 \vee \ldots \vee \neg L_{m-1} \vee \neg L_m)$
- Horn Clause: a clause with a single non-negated literal, e.g:
  $\{\neg L_1 \vee \neg L_2 \vee L_3\} \equiv L_3 \text{:-} L_1, L_2$
- Hypothesis: a set of clauses H
- Completeness: H covers all positive examples
- Consistency: H covers no negative examples

ILP

Approaches

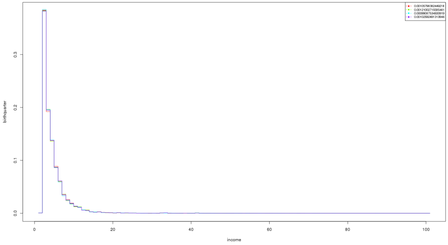
- Bottom-up: Start with least general H then perform generalizations
- Top-down: Start with most general H then perform specializations
  - Specialization loop: adds literals to a clause and ensures consistency
  - Covering loop: adds clauses to the hypothesis and ensures completeness

Correlation between Literals

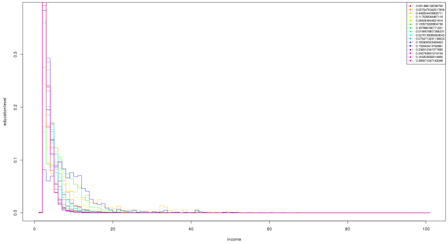
Let's say we want to refine a clause with $hasIncome(x, y)$ with an interval for y. What property is more interesting to add to the clause body:

$$quarterOfBirth(x, z) \text{ or } hasEducation(x, z)?$$

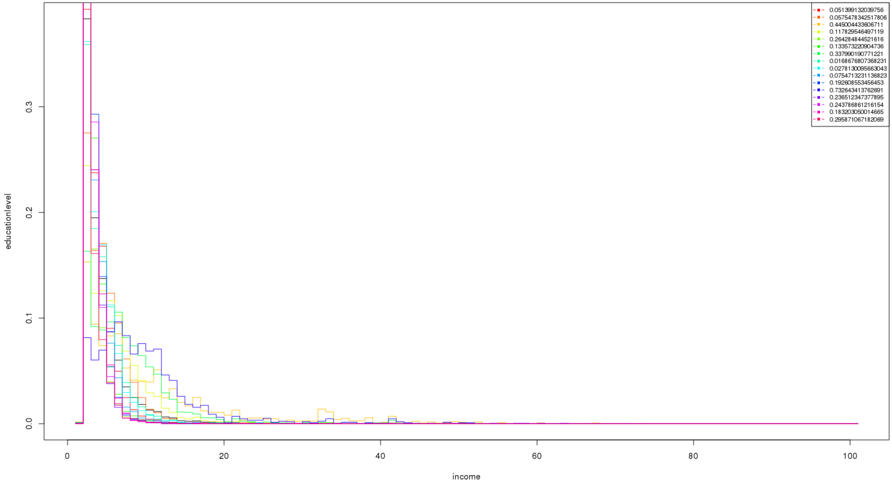
quarterOfBirth



hasEducation



Correlation between Literals



Correlation between Literals

USCensus constants for z in $hasEducation(x, z)$

N/A (less than 3 years old)	High school graduate
No school completed	Some college, less than 1 year
Nursery school to grade 4	One or more years of college, no degree
Grade 5 or grade 6	Associate's degree
Grade 7 or grade 8	Bachelor's degree
Grade 9	Master's degree
Grade 10	Professional school degree
Grade 11	Doctorate degree
Grade 12 no diploma	

Correlation between Literals

Use distribution divergence as interestingness measures, e.g.: Kullback-Leibler, Chi-square, Jensen-Shannon, etc.
But, divergence alone isn't a good idea because:

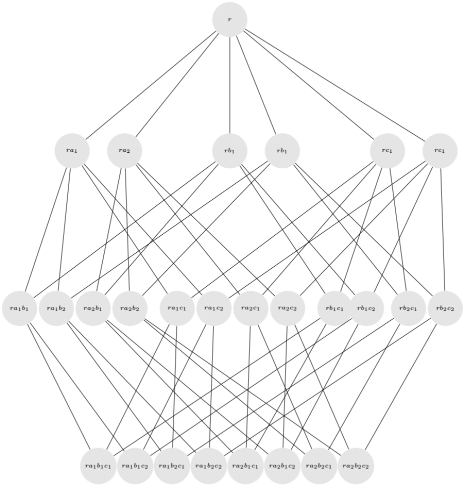
- Lower support histograms are more likely to have a divergent distribution
- Still, support is a good measure as well

Then combine both measures: divergence*support

Correlation Lattice

- Build a lattice similar to an itemset lattice
- Numerical property as root
- The "items" would be literals that can be joined with the root's non-numerical variable
- Each node consists of the joined with a set of literals
- Root's numerical atribute domain is discritezed in k buckets
- Each node x has a histogram $h(x)$ with examples frequencies $h_i(x)$ for each bucket $i \in 1, \ldots, k$ to enable divergence measures
- Then we can use it to suggest the most interesting literals to be added in the refinement step from core-ILP
- Idea is to generate a correlation lattice for each numerical attribute as preprocessing step

Correlation Lattice



$$
\begin{aligned}
r &= hasIncome(x, y) \\
a_1 &= hasSex(x, Male) \\
a_2 &= hasSex(x, Female) \\
b_1 &= employmentStatus(x, Employed) \\
b_2 &= employmentStatus(x, Unemployed) \\
c_1 &= hasDeficiency(x, Yes) \\
c_2 &= hasDeficiency(x, No)
\end{aligned}
$$

Correlation Lattice

- Number of nodes in a lattice with ℓ levels n properties and m constants per property:

$$\sum_{i=1}^{\ell} \binom{nm}{i} \tag{2}$$

- Too expensive, we need to reduce size
  - prune by support (safe)
- If not sufficient, we can restrict the literals to be added in the lattice in order to reduce n and m

Correlation Lattice

Literal Restrictions

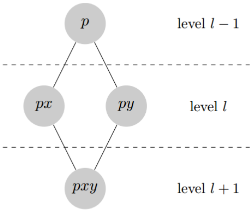
- Lattice literals should directly join with root's non-numerical argument variable
- Other argumets in the literal should either be a free variable or a constant
- Literals that don't directly join with root should be combined with a linking property, e.g.:

  $wasBornIn(x, z) hasOfficialLanguage(z, w)$ as a single literal $r(x, w)$
- This can be used for enable integration with different datasets

$$owl{:}sameAs(x, z) directed(z, w)$$

Independence checks

Checks if a pair of nodes joining nodes are independent given their common parent



- Estimate $\hat{h}(pxy)$ assuming that x and y are independent given p
- Query actual $h(pxy)$ and perform a Pearson's chi-squared test

$$H_0 = x \text{ and } y \text{ are independent given } p$$
$$H_1 = x \text{ and } y \text{ are dependent given } p$$

If there's not enough evidence of dependence, we know that:

$$x\text{:-}py \equiv x\text{:-}p$$
$$y\text{:-}px \equiv y\text{:-}p$$

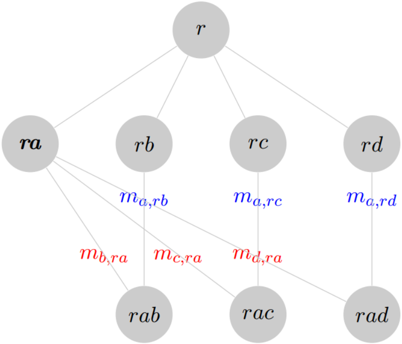
The smaller the *p-value* (or simply the greater the χ^2 value) the greater the evidence that x and y are dependent given p, therefore the more interesting it is to join both py and px

Search in the Lattice

- In the refinement loop from the core-ILP, the clauses have a fixed head and literals are added to the body.
- Assuming that head literal is present in the lattice, we want the interestingness of adding the head to the body.
- Searching for the new literals to be attached to the body that gives you best interestingness when adding the head is not a very simple task.
- e.g., if we have a literal a fixed as head, and wehave the lattice root literal r as body, (i.e., the current clause is a:-r), we want the new literal l such that interestingness of adding a to rl is maximum.

Search in the Lattice

In the following example, we would have b, c, and d as possible new literals



Search in the Lattice

What has to be done?

- Search the node with body literals
- For each child of such node check head literal can be further added, if so collect the new literal and the interestingness value of adding the head
- Sort the possible new literals by interestingness

Alternative?

- Create mapping in every node with the possible head literals as key and sorted literals to be added to body as value, e.g. for the node ra_1b_1 in 18:

$$\begin{array}{c|l} a_1 & c_1 \ [m_{a_1, rb_1c_1}] \\ & c_2 \ [m_{a_1, rb_1c_2}] \\ \hline b_1 & c_1 \ [m_{b_1, ra_1c_1}] \\ & c_2 \ [m_{b_1, ra_1c_2}] \end{array}$$

- Only add entry if head and new literal not independent given body

Experiments

Not done yet! Some experiments done with USCensus

- All data joined by person only (anonymized)
- All properties categorical (categories as literals)
- Not densely linked to other datasets [talk a bit about interestingness measures evaluation done with USCensus]

The End