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:

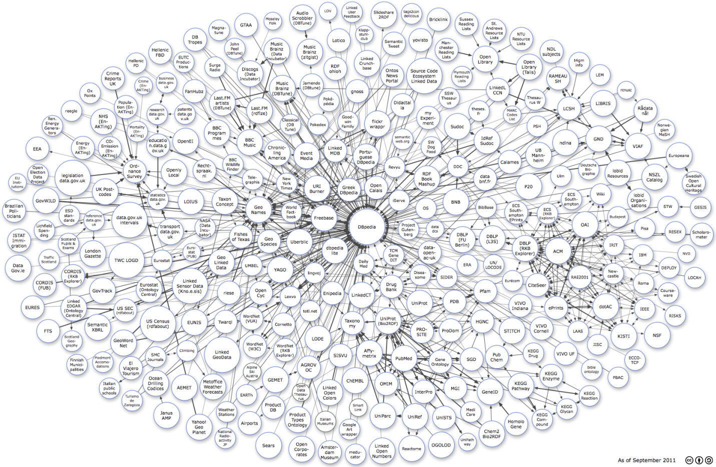
- RDF
- OWL
- SKOS
- SPARQL

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


DBpedia
Freebase
Geo Names
dbpedia lite
UMBEL
YAGO
Uberblic
URI Burner
Linked MDB
Event Media
BBC Music
BBC Program mes
Last.FM (rdfize)
Music Brainz (DBTune)
Audio Scrobbler (DBTune)
GTAA
DB Tropes
Hellenic FBD
Hellenic PD
Crime Reports UK
Ox Points
reegle
EEA
GovWILD
Brazilian Poli- ticians
ISTAT Immi- gration
Data Gov.ie
EURES
FTS
legislation data.gov.uk
data.gov.uk intervals
UK Post- codes
Ord- nance Survey
Openly Local
OpenEI
LOIUS
Taxon Concept
TWC LOGD
Eurostat
London Gazette
GovTrack
US Census (rdfabout)
EUNIS
Linked GeoData
Twarql
LODE
SISVU
PubMed
SGD
GeneID
KEGG Pathway
UniParc
UniRef
UniSTS
OGOLOD
Reactome
Sears
Airports
AEMET
Janus AMP
DBLP (L3S)
ACM
IRIT
IBM
CiteSeer
ePrints
dotAC
KISTI
NSF
IEEE
RISKS
LOCAH
DEPLOY
Roma
Pisa
RESEX
GESIS
STW
NSZL Catalog
Europeana
VIAF
GND
LIBRIS
LCSH
ndlna
Sudoc
IdRef Sudoc
DDC
data bnf.fr
P20
BNB
OAI
Open Library
Bricklink
yovisto
Lotico
gnoss
flickr wrappr
Pokedex
Drug Bank
UniProt
PDB
HGNC
STITCH
VIVO Cornell
Pfam
SIDER
LinkedCT
Enipedia
Lexvo
lingvoj
riese
Open Cyc
As of September 2011

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.:

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

What about numerical attributes?

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

- **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)

ILP

Important concepts:

- Literal: predicate symbol with bracketed n-tuple, e.g:
  $L = livesIn(x, y)$
- Clause: a set of literals (negated or not), e.g:
  $c = \{L_1, L_2, \ldots, \neg L_{m-1}, \neg L_m\}$
- Horn Clause: a clause with a single non-negated literal, e.g:
  $\{L_1, L_2, \neg L_3\} \equiv L_3$:-$L1, L2$
- Hypothesis: a set of clauses H
- Completeness: H covers all positive examples
- Consistency: H covers no negative examples

ILP

Approaches

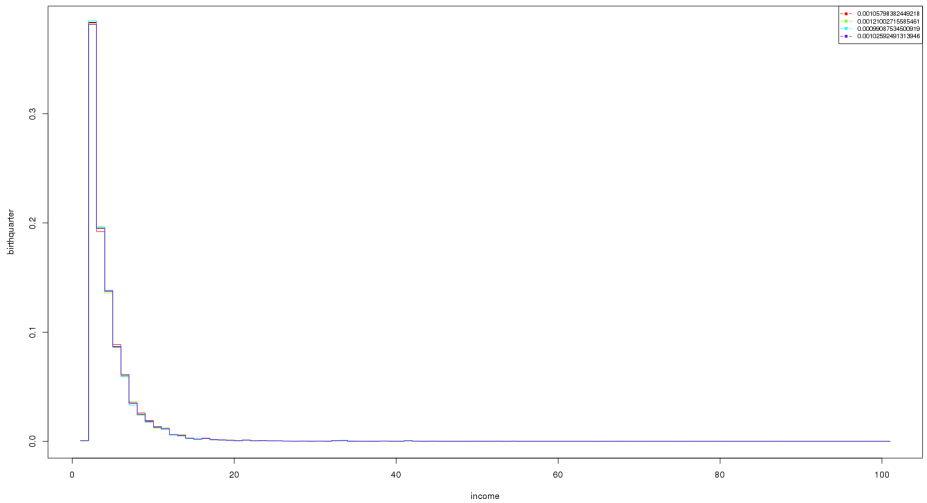
- Bottom-up: Start with least general H then perform generalizations
- Top-down: Start with most general H then perform specializations
  - Specialization loop: adds literals to a clause and ensures consistency
  - Covering loop: adds clauses to the hypothesis and ensures completeness

Correlation between Literals

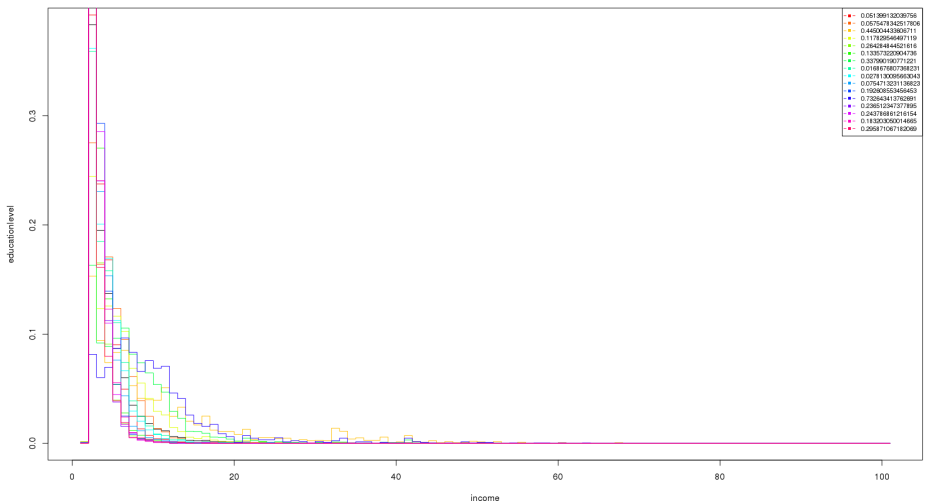
Let's say we want to refine a clause with $hasIncome(x, y)$ with an interval for y. What property is more interesting to add to the clause body:

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

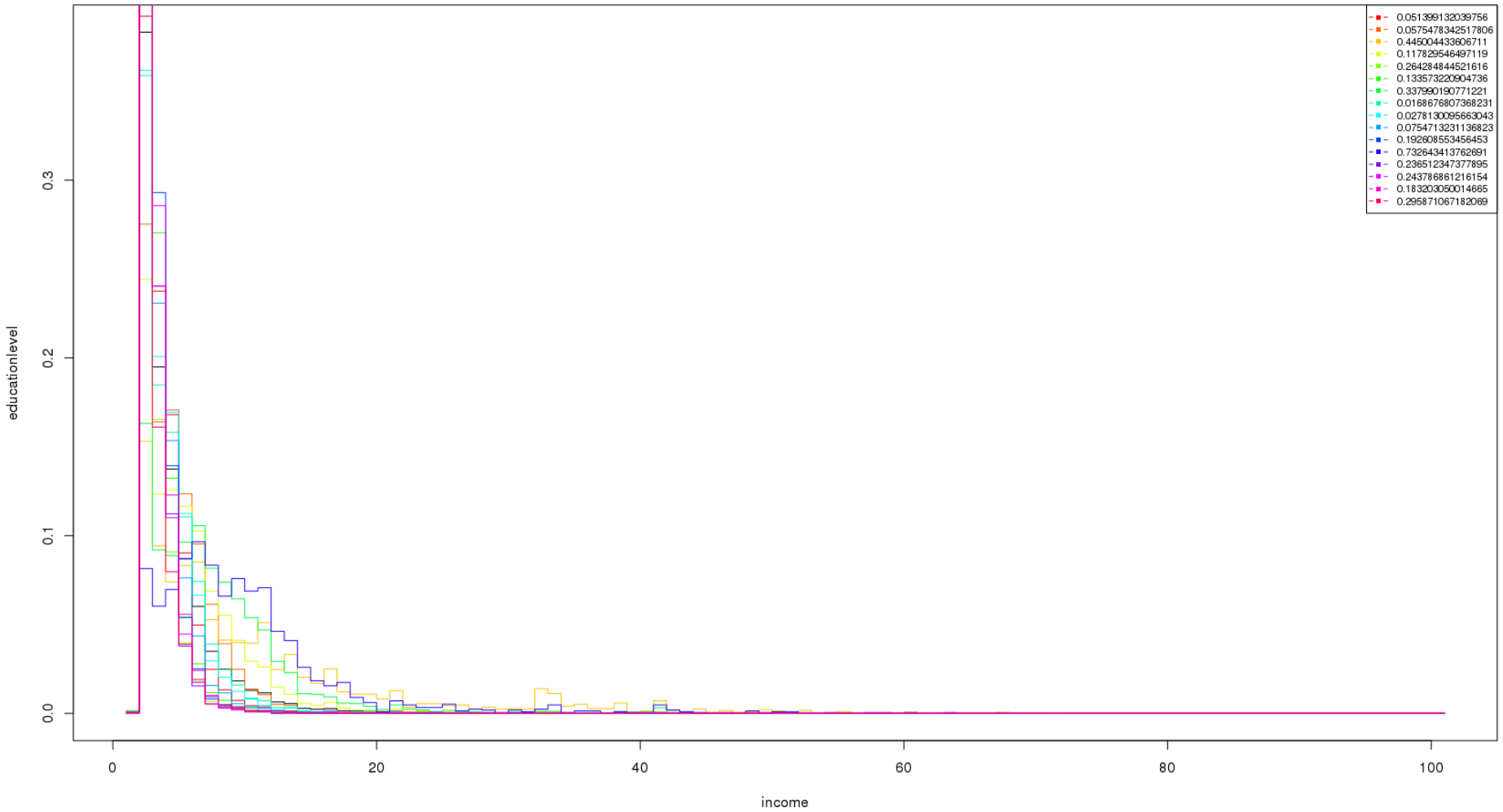
quarterOfBirth



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

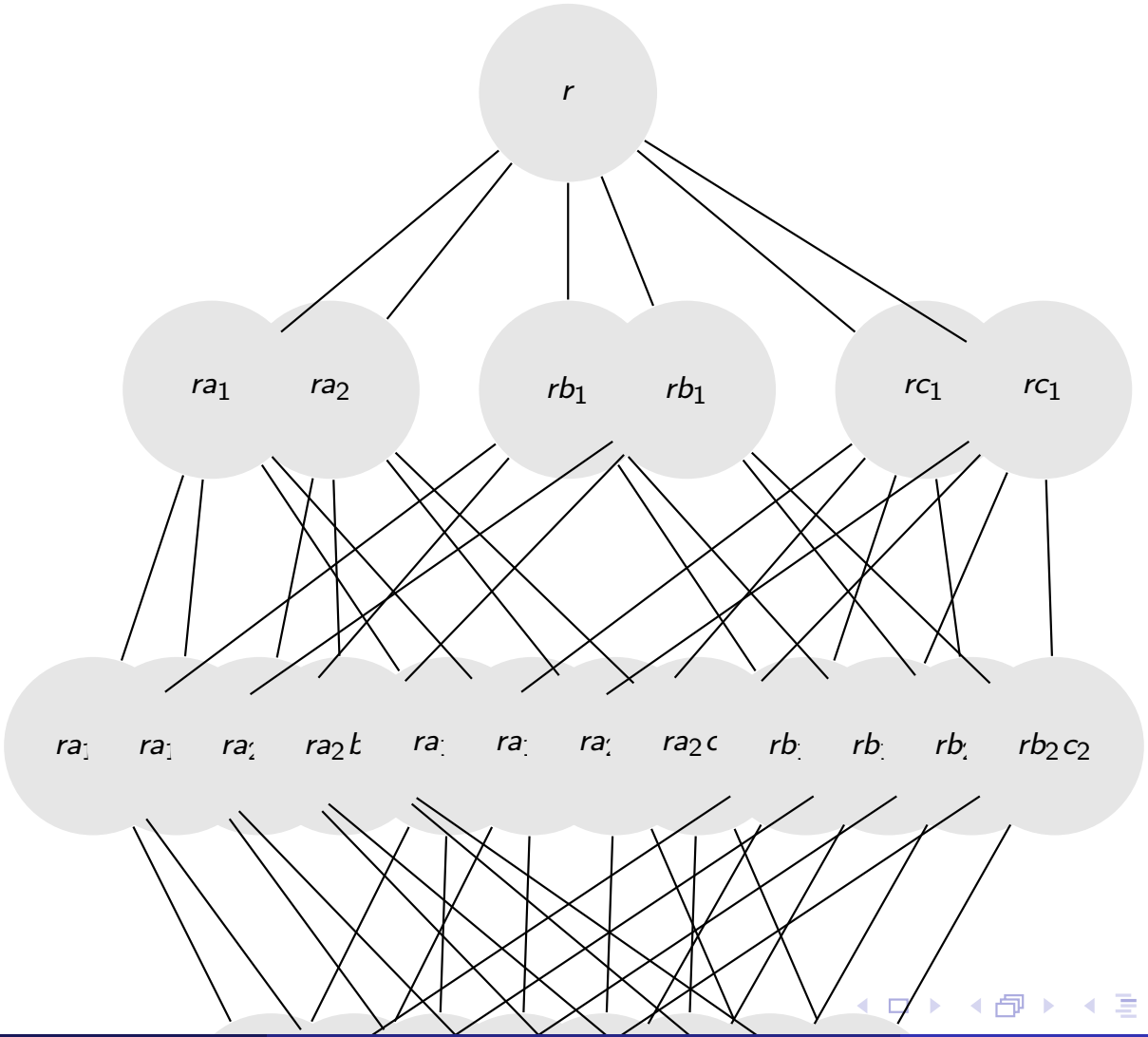
Then combine both measures: divergence*support

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
  Each node has a frequency histogram with examples distribution over root's numerical attribute 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

Figure : Correlation Lattice example



Correlation Lattice

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

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

- 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 as much as possible

Multiple Columns

Heading

1. Statement
2. Explanation
3. Example

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer lectus nisl, ultricies in feugiat rutrum, porttitor sit amet augue. Aliquam ut tortor mauris. Sed volutpat ante purus, quis accumsan dolor.

Table

Treatments	**Response 1**	**Response 2**
Treatment 1	0.0003262	0.562
Treatment 2	0.0015681	0.910
Treatment 3	0.0009271	0.296

Table : Table caption

Theorem

Theorem (Mass–energy equivalence)

$E = mc^2$

Verbatim

Example (Theorem Slide Code)

```
\begin{frame}
\frametitle{Theorem}
\begin{theorem}[Mass--energy equivalence]
$E = mc^2$
\end{theorem}
\end{frame}
```

Figure

Uncomment the code on this slide to include your own image from the same directory as the template .TeX file.

Citation

An example of the `\cite` command to cite within the presentation:

This statement requires citation [Smith, 2012].

References

John Smith (2012)

Title of the publication

Journal Name 12(3), 45 – 678.

The End