Transductive Inference for Class-Membership Propagation in Web Ontologies

Pasquale Minervini Claudia d'Amato Nicola Fanizzi
Floriana Esposito

LACAM, Dipartimento di Informatica, Università degli Studi di Bari "Aldo Moro"
via E. Orabona, 4 - 70125 Bari - Italia
first.last@uniba.it

May 2013

Motivation

Knowledge encoded in Semantic Web (SW) standards is increasing rapidly. Purely deductive inference may have some limitations:

- scalability;
- uncertainty [da Costa et al., 2008];
- uses axiomatic prior knowledge, ignoring statistical regularities;
- requires expensive knowledge engineering processes.

Machine learning can provide scalable solutions to exploit prior knowledge and statistical regularities in data.

Learning from DL Representations

In SW literature, several learning tasks have been faced:

- **Concept learning** ([Lehmann et al., 2010, Fanizzi et al., 2008, Esposito et al., 2004]);
- **Schema** (TBox) induction ([Völker et al., 2011]);
- **Assertion** (ABox) prediction (survey [Rettinger et al., 2012]).

Focus: **assertion prediction** problem

- deciding whether an individual belongs to a given concept;
- often approached using statistical models.

Generative vs. Discriminative Models

Proposed statistical assertion prediction models classified in:

- **Generative models** construct a joint probabilistic model of the variables involved in the prediction task;
- **Discriminative models** only focus on the actual accuracy of the induced predictive model.

Building a generative prediction models can possibly be an *harder* problem than building a discriminative one.

Generative vs. Discriminative Models

Real-life ontologies characterized by abundance of unlabelled instances and few labelled ones:

- Unknown concept-membership relations considered as *unlabelled instances*.

Generative models can be learned with unlabelled instances

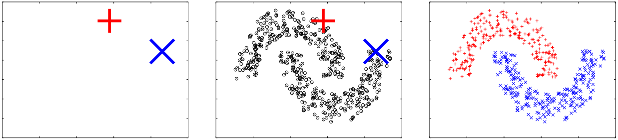
- By maximizing the likelihood of observable data, e.g. by means of the classic EM framework.

How to build discriminative models accounting for unlabelled instances ?

Leveraging unlabelled instances

Proposal: adoption of *Semi-Supervised Learning* (SSL) for building discriminative models from SW representations.

- **Semi-Supervised Smoothness Assumption** – if two points are linked by a path through a high-density area, they are likely to be associated to closely located labels.



(a) -Unlabelled (b) +Unlabelled (c) Actual labels

Figure: The distribution of instances can be informative wrt. the distribution of labels

Inductive vs. Transductive Semi-Supervised Learning

- **Inductive SSL** – learn a prediction function expected to be a good predictor on future, unseen instances;
- **Transductive SSL** – learn a prediction function expected to be a good predictor on available, unlabeled instances.

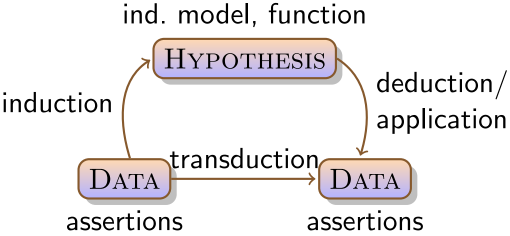
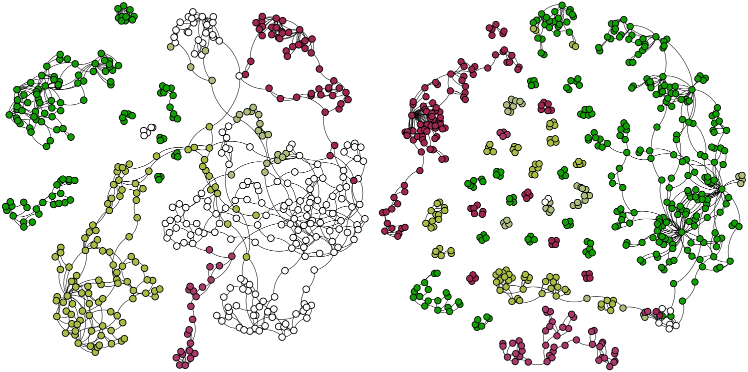


Figure: Transductive and inductive inference

Performing Transductive Semi-Supervised inference

- Looking for a *smooth* labelling is a *simpler problem* than finding a smooth prediction function wrt. the instance space.
- We will thus focus on *Transductive Semi-Supervised inference*:
  - **Aim:** given a set of labelled and unlabelled instances, find a labelling consistent with labelled instances that *varies smoothly* among similar instances.
  - **Proposed approach:** *Graph-based regularization* (consisting in *penalizing non-smooth labellings*) allows to:
    - Scale up to large-size sets of instances;
    - Provide confidence indicators about inferred knowledge.
- Algorithm:
  - Create a k-NN *similarity graph*: each individual linked with top k similar individuals ($\{0, 1\}$ adjacency matrix $\mathbf{W}$);
  - Find labels minimizing a cost function enforcing consistency (wrt. training labels) and smoothness (wrt. the graph).
- Similarity relations calculated using *graph* and *DL kernels*.

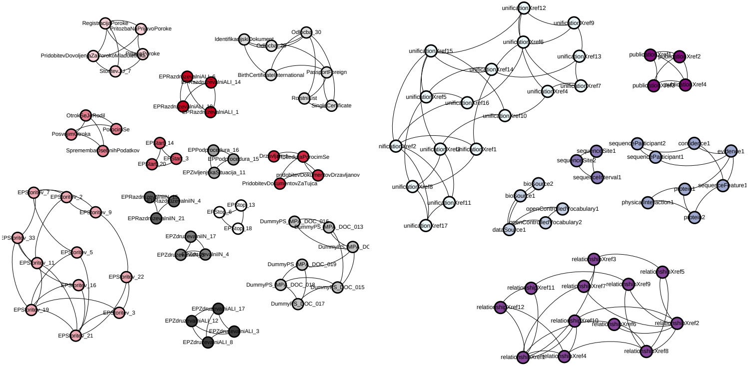
Clustered structures in Web Ontologies



(a) AIFB Affiliations (Persons) (b) AIFB Affiliations (Articles)

Figure: Neighbourhood relations (wrt. the SubTree kernel [Lösch et al., 2012]) in the AIFB Affiliations ontology

Clustered structures in Web Ontologies

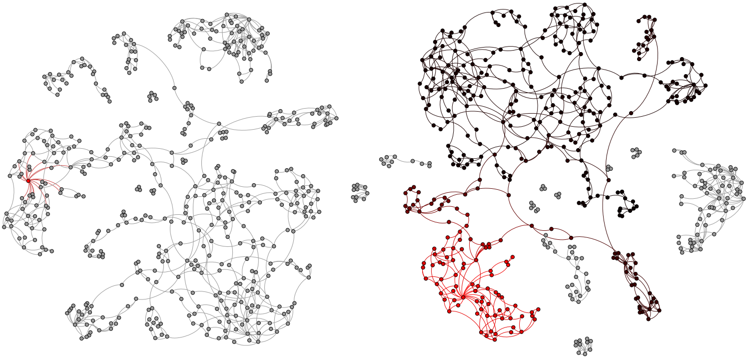


(a) Leo

(b) BioPAX

Figure: Neighbourhood relations (wrt. a DL kernel) between individuals in two ontologies from the TONES repository

Propagating knowledge in the AIFB Affiliations ontology



(a) AIFB Affiliations (Persons) (b) AIFB Affiliations (Articles)

Figure: Example of label propagation via graph-based regularization in the AIFB Affiliations ontology.

Propagating membership knowledge among instances

Assume:

- predicted labellings encoded in a vector $\hat{\mathbf{f}} \in (-1,1)^N$;
- similarity relations encoded in a symmetric $\{0,1\}$ adjacency matrix $\mathbf{W}$.

Need a *cost function* that both:

1. Enforces consistency with training examples;
2. Enforces smoothness among similar individuals.

Many options to define such costs. By adopting *quadratic cost criteria*, two options were considered:

- $C(\hat{\mathbf{f}}) = ||\hat{\mathbf{f}}_L - \mathbf{f}_L||^2 + \mu \hat{\mathbf{f}}^T \mathbf{L} \hat{\mathbf{f}} + \mu\epsilon ||\hat{\mathbf{f}}||^2$;
  - note that $\hat{\mathbf{f}}^T \mathbf{L} \hat{\mathbf{f}} = 0.5 \sum_{i,j} W_{ij} (\hat{f}_i - \hat{f}_j)^2$;
- $C'(\hat{\mathbf{f}}) = ||\hat{\mathbf{f}} - \mathbf{f}||^2 + \mu \hat{\mathbf{f}}^T \mathcal{L} \hat{\mathbf{f}}$;

where:

- $\mathbf{L} = (\mathbf{D} - \mathbf{W})$ *graph Laplacian* ($\mathbf{D}$ diagonal: $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$);
- $\mathcal{L} = \mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2}$ *normalized graph Laplacian*.

Propagating membership knowledge among instances

Finding $\mathbf{f}^* = \arg\min_{\mathbf{f}} C(\mathbf{f})$ reduces to solving a (large) sparse linear system; e.g. let $C(\hat{\mathbf{f}}) = ||\hat{\mathbf{f}}_L - \mathbf{f}_L||^2 + \hat{\mathbf{f}}^T\mathbf{L}\hat{\mathbf{f}} + \mu\epsilon||\hat{\mathbf{f}}||^2$:

$$\begin{aligned}
\frac{1}{2}\frac{\partial C(\hat{\mathbf{f}})}{\partial \hat{\mathbf{f}}} &= \mathbf{S}(\hat{\mathbf{f}} - \mathbf{f}) + \mu\mathbf{L}\hat{\mathbf{f}} + \mu\epsilon\hat{\mathbf{f}}; \\
\arg\min_{\hat{\mathbf{f}}} C(\hat{\mathbf{f}}) &= (\mathbf{S} + \mu\mathbf{L} + \mu\epsilon\mathbf{I})^{-1}\mathbf{S}\mathbf{f}; \\
\rightarrow x &= A^{-1}b;
\end{aligned}$$

Solving a sparse linear system is a well-known problem whose time complexity is nearly linear in the number of non-zero entries in the coefficient matrix [Spielman and Teng, 2004].

Working around the irrelevant roles issue

Proposed approach: combine several *heterogeneous* similarity/affinity graphs (encoding e.g. co-authorship, friendship, collaboration relations).
Different graph-based regularizers combined together:

- **HMRW**: regularizer $\mu \hat{\mathbf{f}}^T \mathbf{L} \hat{\mathbf{f}}$ becomes $\sum_{r=1}^{R} \mu_r \hat{\mathbf{f}}^T \mathbf{L}_r \hat{\mathbf{f}}$;

where $\mathbf{L}_r$ is the graph Laplacian of the r-th similarity/affinity graph, and $\mu_r \geq 0$ is a weight associated to each of such graphs.

- Evaluation setting: predicting research group affiliations in the AIFB Affiliations ontology [Lösch et al., 2012];
- Evaluated approaches:
  - **RG**: $C(\hat{\mathbf{f}}) = ||\hat{\mathbf{f}}_L - \mathbf{f}_L||^2 + \mu \hat{\mathbf{f}}^T \mathbf{L} \hat{\mathbf{f}} + \mu\epsilon ||\hat{\mathbf{f}}||^2$;
  - **MRW**: like C, but $\mu \to 0$ (enforced training labels);
  - **HRMW**: linear comb. of multiple graph-based regularizers;
  - **SVM**: state of the art discriminative method.
  - C' defined previously found too conservative.

Results

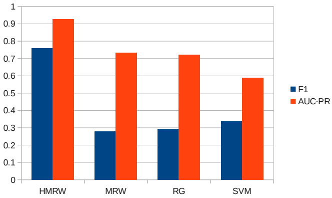
Method	F1	AUC-PR
HMRW	**0.76 ± 0.207**	**0.928 ± 0.117**
MRW	0.278 ± 0.202	0.734 ± 0.166
RG	0.293 ± 0.218	0.723 ± 0.166
SVM	0.34 ± 0.177	0.588 ± 0.16

(a) Training: 9/10 folds

Method	F1	AUC-PR
HMRW	**0.738 ± 0.118**	**0.907 ± 0.073**
MRW	0.276 ± 0.163	0.665 ± 0.095
RG	0.271 ± 0.17	0.665 ± 0.111
SVM	0.338 ± 0.167	0.488 ± 0.092

(b) Training: 5/10 folds

Figure: 10-fold CV results for SVM, RG, MRW and HMRW



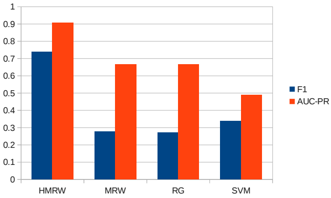


Figure: 10-fold CV results using 9 and 5 training folds

Conclusions

- Tackling the problem of (semi-supervised) transductive inference from SW representations.
  - source: `code.google.com/p/knowledge-propagation/`
- From evaluations, it emerged that single graph (and DL) kernels may tend to be domain dependant:
  - Sometimes they work well, sometimes they do not;
  - It gives hints of how well a kernel may suit a problem.
- Current effort: learning a task-dependent combination of kernels, taking in account different aspects of the domain (from terminology to relational structure);

da Costa, P. C. G. et al., editors (2008).
Uncertainty Reasoning for the Semantic Web I, ISWC International Workshops, URSW 2005-2007, Revised Selected and Invited Papers, volume 5327 of *Lecture Notes in Computer Science*.
Springer.

Esposito, F., Fanizzi, N., Iannone, L., Palmisano, I., and Semeraro, G. (2004).
Knowledge-intensive induction of terminologies from metadata.
In McIlraith, S. A., Plexousakis, D., and van Harmelen, F., editors, *International Semantic Web Conference*, volume 3298 of *Lecture Notes in Computer Science*, pages 441–455. Springer.

Fanizzi, N., d'Amato, C., and Esposito, F. (2008).
Dl-foil concept learning in description logics.
In Zelezný, F. and Lavrac, N., editors, *ILP*, volume 5194 of *Lecture Notes in Computer Science*, pages 107–121. Springer.

Lehmann, J. et al. (2010).
Concept learning in description logics using refinement operators.
Machine Learning, 78(1-2):203–250.

Lösch, U., Bloehdorn, S., and Rettinger, A. (2012).
Graph kernels for rdf data.
In Simperl, E. et al., editors, *ESWC*, volume 7295 of *Lecture Notes in Computer Science*, pages 134–148. Springer.

Rettinger, A., Lösch, U., Tresp, V., d'Amato, C., and Fanizzi, N. (2012).
Mining the Semantic Web: Statistical learning for next generation knowledge bases.
Data Min. Knowl. Discov., 24(3):613–662.

Spielman, D. A. and Teng, S.-H. (2004).
Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems.
In *Proceedings of the 36th ACM Symposium on Theory of Computing, STOC'04*, pages 81–90. ACM.

Völker, J. et al. (2011).
Statistical schema induction.
In Antoniou, G., Grobelnik, M., Simperl, E. P. B., Parsia, B., Plexousakis, D., Leenheer, P. D., and Pan, J. Z., editors, *ESWC (1)*, volume 6643 of *Lecture Notes in Computer Science*, pages 124–138. Springer.