A Graph Regularization Based Approach to Transductive Class-Membership Prediction

Pasquale Minervini, Claudia d'Amato, Nicola Fanizzi

Computer Science Department • *University of Bari, Italy*

November, 11, 2012 - Boston

Contents

Introduction

- Deductive Reasoning, usually adopted in the SW context, may fail in presence of inconsistent and/or noisy knowledge bases
- Machine learning methods can be adopted to *perform approximate and uncertain reasoning*
    - allowing to derive *conclusions which are not derivable* or refutable from the knowledge base
- **Issue:** unlabeled instances could be present (because of the OWA)

Focus

- **Goal:** Class-membership (concept) prediction task
- **Focus on** *trasductive semi-supervised learning methods*
    - family of machine learning methods that *use both labeled and unlabeled data* for training a learning algorithm
    - in trasductive setting, the learning algorithm only aims at estimating the class-membership for a given training set, without generalizing to individuals outside such set.
- **Motivation:** *automation of the knowledge acquisition process*
    - the acquisition of labeled (training) data for a learning task often requires the manual effort of human agents ⇒ *the cost may render* a fully labeled training set *infeasible*
    - the acquisition of unlabeled data is relatively inexpensive
    - labelled data are not alway available

Transductive Class-Membership Prediction

Definition

Given:

- a *target* concept C;
- a set of training individuals $\mathsf{Ind}_C(\mathcal{K})$ in a knowledge base $\mathcal{K}$ partitioned in
  - $\mathsf{Ind}^+_C(\mathcal{K}) = \{a \in \mathsf{Ind}_C(\mathcal{K}) \mid \mathcal{K} \models C(a)\}$ positive examples,
  - $\mathsf{Ind}^-_C(\mathcal{K}) = \{a \in \mathsf{Ind}_C(\mathcal{K}) \mid \mathcal{K} \models \neg C(a)\}$ negative examples,
  - $\mathsf{Ind}^0_C(\mathcal{K}) = \{a \in \mathsf{Ind}_C(\mathcal{K}) \mid \mathcal{K} \not\models C(a) \wedge \mathcal{K} \not\models \neg C(a)\}$ unlabeled examples;
- A *cost function* $cost(\cdot) : \mathcal{F} \mapsto \mathbb{R}$, specifying the *cost* associated to a set of class-memberships assigned to training individuals by $f \in \mathcal{F}$, where $\mathcal{F}$ is a space of labelling functions of the form $f : \mathsf{Ind}_C(\mathcal{K}) \mapsto \{+1, -1\}$;

Find: a *labelling function* $f^* \in \mathcal{F}$ **minimizing** the given cost function w.r.t. $\mathsf{Ind}_C(\mathcal{K})$:

$$f^* \leftarrow \arg\min_{f \in \mathcal{F}} cost(f).$$

The function f^* can then be used to estimate the class-membership w.r.t. C for all training individuals $a \in \mathsf{Ind}_C(\mathcal{K})$

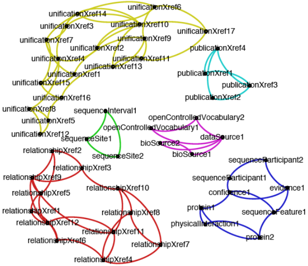
Graph-based semi-supervised approach

1. Choose/build a *target concept C*
2. Determine the training set $\mathrm{Ind}_C(\mathcal{K})$ w.r.t. C in $\mathcal{K}$ as given by positive, negative and unlabeled instances
3. Build the Nearest Neighbor (NN) Semantic Similarity graph
4. Define a cost over functions $f \in \mathcal{F} \Rightarrow$ *quadratic cost criterion framework* as a cost function
    - *finding a labeling function that is*
        - consistent with the given labels $\Rightarrow$ loss function as a measure of consistency with the given labels
        - changes smoothly between similar instances $\Rightarrow$ *Regularization by graph* $\Rightarrow$ measure of smoothness among the similarity graph as a regularizer

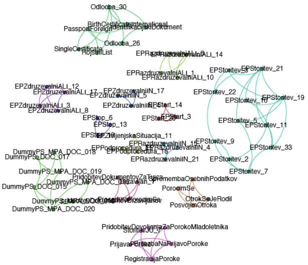
Building the NN-Semantic Similarity graph

The *Similarity graph* is built as a matrix $\mathbf{W}$ where $\mathbf{W}_{ij}$ is the similarity value between two training examples x_i and x_j

- a NN graph, for each instance x_i, contains similarity the value only for the k most similar instances (the others are set to 0)
- employed a family of similarity measures between in individuals in a DL knowledge base *[d'Amato et al. @ URSW'09]*



(a) 3-NN graph, BioPAX (Proteomics)



(b) 3-NN graph, Leo

The Family of Similarity Measure

Let $\mathcal{K} = (\mathcal{T}, \mathcal{A})$ be a DL knowledge base. Given a set of concept descriptions $\mathsf{F} = \{F_1, F_2, \ldots, F_m\}$, corresponding weights $w_1, \ldots, w_m$, and $p > 0$, a family of dissimilarity functions $d_p^{\mathsf{F}} : \mathrm{Ind}(\mathcal{A}) \times \mathrm{Ind}(\mathcal{A}) \mapsto [0, 1]$ is defined by:

$$\forall a, b \in \mathrm{Ind}(\mathcal{A}): \quad d_p^{\mathsf{F}}(a, b) := \frac{1}{|\mathsf{F}|} \left[\sum_{i=1}^{|\mathsf{F}|} w_i \mid \delta_i(a, b) \mid^p \right]^{1/p},$$

where the dissimilarity function δ_i ($i \in \{1, \ldots, m\}$) is defined by:

$$\forall a, b \in \mathrm{Ind}(\mathcal{A}): \quad \delta_i(a, b) = \begin{cases} 0 & F_i(a) \in \mathcal{A} \wedge F_i(b) \in \mathcal{A} \\ 1 & F_i(a) \in \mathcal{A} \wedge \neg F_i(b) \in \mathcal{A} \text{ or} \\ & \neg F_i(a) \in \mathcal{A} \wedge F_i(b) \in \mathcal{A} \\ 1/2 & \mathit{otherwise}. \end{cases}$$

Quadratic Cost Criteria: Preliminaries

1. Original label space $\{-1, +1\}$ relaxed to $[-1, +1]$ $\Rightarrow$ allows to express the confidence associated to a labeling
   - The labeling function space $\mathcal{F}$ is relaxed to functions of the form $f : \mathrm{Ind}_C(\mathcal{K}) \mapsto [-1, +1]$
   - labeling functions can be represented as vectors $\mathbf{y} \in [-1, +1]^n$ where $n = |\mathrm{Ind}_C(\mathcal{K})|$
2. Let $\hat{\mathbf{y}} \in [-1, +1]^n$ be a possible labeling for n instances
   - $\hat{\mathbf{y}}$ be seen as a $(l + u) = n$ dimensional vector, where the first l indices refer to labeled instances, and the last u to unlabeled instances: $\hat{\mathbf{y}} = [\hat{\mathbf{y}}_l, \hat{\mathbf{y}}_u]$

Quadratic Cost Criteria...

- **Consistency** of $\hat{\mathbf{y}}$ w.r.t. the original labels can be formulated *in the form of a quadratic cost*

$$\sum_{i=1}^{l}(\hat{y}_i - y_i)^2 = ||\hat{\mathbf{y}}_l - \mathbf{y}_l||^2$$

- labellings can be regularised w.r.t. the graph structure *alternatively as*:
  1. $$0.5 \sum_{i,j=1}^{n} \mathbf{W}_{ij}(\hat{y}_i - \hat{y}_j)^2 = \hat{\mathbf{y}}^T \mathbf{L} \hat{\mathbf{y}}$$

     where $\mathbf{W}$ is the semantic similarity graph and $\mathbf{L} = \mathbf{D} - \mathbf{W}$, with D the diagonal matrix s.t. $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$, is the unnormalized graph Laplacian **[Belkin et al. @ COLT'04]**
  2. $(\mathbf{D}^{-0.5}\hat{\mathbf{y}})^T \mathbf{L} (\mathbf{D}^{-0.5}\hat{\mathbf{y}})$ **[Zhou et al. @ ICML'5]**

...Quadratic Cost Criteria

Putting consistency and regularization together two *quadratic cost criteria* are obtained:

1. **Regression on Graph (RG)RG:**
   $cost(\hat{\mathbf{y}}) = ||\hat{\mathbf{y}}_l - \mathbf{y}_l||^2 + \mu\hat{\mathbf{y}}^T \mathbf{L}\hat{\mathbf{y}} + \mu\epsilon||\hat{\mathbf{y}}||^2$; **[Belkin et al. @ COLT'04]**
2. **Consistency Method (CM) CM:**
   $cost(\hat{\mathbf{y}}) = ||\hat{\mathbf{y}}_l - \mathbf{y}_l||^2 + \mu(\mathbf{D}^{-0.5}\hat{\mathbf{y}})^T \mathbf{L}(\mathbf{D}^{-0.5}\hat{\mathbf{y}}) + ||\hat{\mathbf{y}}_u||^2$. **[Zhou et al. @ ICML'5]**

By recurring to derivative, finding the minimum for **RG** (resp. **CM**) consists in solving a (possibly sparse) linear system whose *time complexity is* nearly *linear* in the number of non-zero entries in the coefficient matrix

Experiments: Setup

GOAL: evaluation and comparison of *inductive and trasductive methods* for class-memebership prediction

- 20 random queries C created for each ontology (Tab.)
  - C (resp. $\neg C$) should contain at least 10 instances
  - Pellet Reasoner v2.3.0 emploied for building the *TrSet* $\mathrm{Ind}_C(\mathcal{K})$
- ten-fold cross validation adopted
- *Trasductive Methods:* Regression Graph (**RG**), Consistency Method (**CM**), Label Propagation ((**LP**))
- *Inductive Methods:* Soft Margin SVM (**SM-SVM**), Laplacian SVM (**LapSVM**), k-NN ($k = \sqrt{l}$ where $l =$ num. labeled ex. (($\sqrt{l}$-**NN**))

Ontology	Expressivity	#Axioms	#Indiv.	#Classes	#Obj.Prop.
BioPax (Prot.)	$\mathcal{ALCHN(D)}$	773	49	55	47
Family-Tree	$\mathcal{SROIF(D)}$	2059	368	22	52
Leo	$\mathcal{ALCHIF(D)}$	430	61	32	26
MDM0.73	$\mathcal{ALCHOF(D)}$	1098	112	196	22
Wine	$\mathcal{SHOIN(D)}$	1046	218	142	21

Metrics

Match Rate Case of an individual that got the same label by the reasoner and the inductive classifier.

Omission Error Case of an individual for which the inductive method could not determine whether it was relevant to the query concept or not while it was found relevant by the reasoner.

Commission Error Case of an individual found to be relevant to the query concept while it logically belongs to its negation or vice-versa.

Induction Case of an individual found to be relevant to the query concept or to its negation, while either case is not logically derivable from the knowledge base.

Results

Leo	Match	Omission	Commission	Induction
RG	1 ± 0	0 ± 0	0 ± 0	0 ± 0
CM	1 ± 0	0 ± 0	0 ± 0	0 ± 0
LP	0.942 ± 0.099	0.007 ± 0.047	0.052 ± 0.091	0 ± 0
SM-SVM	0.963 ± 0.1	0 ± 0	0.037 ± 0.1	0 ± 0
LapSVM	0.978 ± 0.068	0 ± 0	0.022 ± 0.068	0 ± 0
$\sqrt{l}$-NN	0.971 ± 0.063	0 ± 0	0.029 ± 0.063	0 ± 0

BioPAX (Prot.)	Match	Omission	Commission	Induction
RG	0.986 ± 0.051	0.004 ± 0.028	0.008 ± 0.039	0.002 ± 0.02
CM	0.986 ± 0.051	0.002 ± 0.02	0.01 ± 0.044	0.002 ± 0.02
LP	0.982 ± 0.058	0.002 ± 0.02	0.014 ± 0.051	0.002 ± 0.02
SM-SVM	0.972 ± 0.075	0 ± 0	0.026 ± 0.068	0.002 ± 0.02
LapSVM	0.972 ± 0.075	0 ± 0	0.026 ± 0.068	0.002 ± 0.02
$\sqrt{l}$-NN	0.972 ± 0.075	0 ± 0	0.026 ± 0.068	0.002 ± 0.02

MDM0.73	Match	Omission	Commission	Induction
RG	0.953 ± 0.063	0.003 ± 0.016	0.011 ± 0.032	0.015 ± 0.039
CM	0.953 ± 0.063	0.001 ± 0.009	0.013 ± 0.036	0.018 ± 0.04
LP	0.942 ± 0.065	0 ± 0	0.026 ± 0.046	0.033 ± 0.054
SM-SVM	0.793 ± 0.252	0 ± 0	0.174 ± 0.255	0.033 ± 0.054
LapSVM	0.915 ± 0.086	0 ± 0	0.052 ± 0.065	0.033 ± 0.054
$\sqrt{l}$-NN	0.944 ± 0.069	0 ± 0	0.023 ± 0.051	0.033 ± 0.054

Wine	Match	Omission	Commission	Induction
RG	0.24 ± 0.03	0 ± 0.005	0.007 ± 0.017	0.5 ± 0.176
CM	0.242 ± 0.028	0 ± 0.005	0.005 ± 0.015	0.326 ± 0.121
LP	0.239 ± 0.035	0 ± 0.005	0.008 ± 0.021	0.656 ± 0.142
SM-SVM	0.235 ± 0.036	0 ± 0	0.012 ± 0.024	0.753 ± 0.024
LapSVM	0.238 ± 0.033	0 ± 0	0.009 ± 0.021	0.753 ± 0.024
$\sqrt{l}$-NN	0.241 ± 0.031	0 ± 0	0.006 ± 0.018	0.753 ± 0.024

Discussion

- FAMILY-TREE ontology (not reported), provided 0.76 ± 0.13 match rates and 0.24 ± 0.13 induction rates for all but **LP** method, where the induction rates were 0.21 ± 0.14
- In general, LapSVM outperformed the other two non-SSL SVM classification methods
- Trasductive approaches generally outperform inductive approaches in terms of *commission error* and *match rate*
- Trasductive approaches resulted more conservative than inductive approaches for MDMO.73 and WINE ontologies, showing
  - highest omission rates
  - lowest induction rates
- The proposed **RG** and **CM** always outperform the **LP** adopted as a baseline trasductive methods

Conclusions

Conclusions: A method for trasductive class-membership prediction based on graph-based regularization has been proposed

- it relies on quadratic cost criteria whose optimization can be reduced to solve a (possibly sparse) linear system
  - nearly linear time complexity in the number of non-zero entries in the coefficient matrix
- Experimental evaluations showed the improvement of the performance of the trasductive approach over the inductive one particularly in terms of commission error and match rate

Future Works:

- Deeply investigate on the correlation between the order of magnitude of unlabeled instances and the results of the proposed method

That's all!

Questions?