On Discovering Deterministic Relationships in Multi-Label Learning via Linked Open Data

Eirini Papagiannopoulou
Grigorios Tsoumakas
Nick Bassiliades

Department of Informatics
Aristotle University of Thessaloniki, Greece
Overview

- What: discovering deterministic relationships among labels
  - Common parents
  - Mutual exclusion
  - Parent-child relationship among existing labels
- How: Using Linked Open Data
- Why:
  - Improve prediction accuracy
  - Reduce inference complexity
**Multilabel Learning**

- Multi-label Learning:

<table>
<thead>
<tr>
<th>attribute 1</th>
<th>attribute 2</th>
<th>...</th>
<th>attribute n</th>
<th>label 1</th>
<th>label 2</th>
<th>...</th>
<th>label k</th>
</tr>
</thead>
<tbody>
<tr>
<td>service</td>
<td>statistics</td>
<td>...</td>
<td>java</td>
<td>TAG_academic</td>
<td>TAG_ajax</td>
<td>...</td>
<td>TAG_api</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>...</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>...</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>
Our Approach – Step 1

• Tokenization of label names

  • In textual datasets names of labels may coincide with names of attributes

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Attribute name</th>
<th>Label Name</th>
<th>Final label name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibtex</td>
<td>model</td>
<td>model</td>
<td>TAG_model</td>
</tr>
<tr>
<td>IC2011</td>
<td>autumn</td>
<td>autumn</td>
<td>autumn_attr</td>
</tr>
</tbody>
</table>

• Tokenize with a standard set of delimiters: `- _ "" , .`

• Tokens that appear in all label names are removed

• The rest of them are concatenated again with a space

Landscape_Nature_attr → Landscape, Nature, attr → Landscape Nature
Our Approach – Step 2

• Look up each label name in WordNet.

→ multiple senses: the correct sense is the most frequent one according to WordNet (based on semantically tagged corpora)

Winter

- the coldest season of the year (24)
- spend the winter (2)

Summer

- the warmest season of the year (58)
- the period of nest development, happiness, or beauty (0)
- spend the summer (0)
Our Approach – Step 3

- Take recursively hypernym-synsets of the determined sense up to the root of WordNet
Our Approach – Step 4

• Examine all pairs of labels (found in WordNet) → their common ancestors are used for expanding the original label space

• Ignore some common parents that appear at the top of the WordNet hierarchy, because they will be added for all label pairs (e.g. entity, abstraction etc)
Our Approach – Step 4

- entity
  - abstraction
    - Measure:quantity:amount
      - period
        - season
          - Winter:wintertime

- entity
  - abstraction
    - Measure:quantity:amount
      - period
        - season
          - Summer:summertime
Our Approach – Step 5

For every new label of each instance:

• If at least one of its children is true then the specific new label will be true

• In all other cases the new label will be false for the specific instance

• In our example:

If (Winter = true or Summer = true)
    Season ← true
Otherwise
    Season ← false
Experiments – Set up

- Calibrated Label Ranking (CLR) – Linear Support Vector Machines
- Mulan\(^1\) and Weka
- 6 multi-label datasets: their labels do not have obscure names i.e. class1, class2 etc
- Split each dataset in train set and test set → 70% - 30%
- Mean Average Precision (MAP) and Logarithmic Loss (LL)

\[^1\) http://mulan.sourceforge.net\]
Two versions of Our Approach

• LOD1
  • Hypernym-synsets of the determined sense of the label for up to two layers
  • We ignore the following 32 senses:
    
    *substance, content, message, theme, topic, subject, domain, activity, individual, someone, somebody, mortal, soul, organism, being, cause, go, locomote, formation, alter, modify, change, alteration, modication, happening, occurence, occurent, whole, unit, object, entity, abstraction*

• LOD2
  • Hypernym-synsets of the determined sense of the label for all layers up to the root of WordNet
  • We ignore 5 senses shown above
## Experimental Results (1)

<table>
<thead>
<tr>
<th>dataset</th>
<th>LOD1</th>
<th>LOD2</th>
<th>%found</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>IC2011</td>
<td>17</td>
<td>69</td>
<td>75 / 99</td>
<td>8000</td>
</tr>
<tr>
<td>bibtex</td>
<td>8</td>
<td>63</td>
<td>119 / 159</td>
<td>7395</td>
</tr>
<tr>
<td>delicious</td>
<td>190</td>
<td>319</td>
<td>778 / 983</td>
<td>16105</td>
</tr>
<tr>
<td>bookmarks</td>
<td>22</td>
<td>79</td>
<td>148 / 208</td>
<td>87856</td>
</tr>
<tr>
<td>corel5k</td>
<td>84</td>
<td>214</td>
<td>367 / 374</td>
<td>5000</td>
</tr>
<tr>
<td>IMDB-F</td>
<td>5</td>
<td>14</td>
<td>23 / 28</td>
<td>120919</td>
</tr>
</tbody>
</table>
# Experimental Results (2)

<table>
<thead>
<tr>
<th>dataset</th>
<th>MAP</th>
<th></th>
<th></th>
<th>Log-Loss</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CLR</td>
<td>LOD1</td>
<td>LOD2</td>
<td>CLR</td>
<td>LOD1</td>
<td>LOD2</td>
</tr>
<tr>
<td>IC2011</td>
<td>.3275</td>
<td>.3263</td>
<td>.3245</td>
<td>.7221</td>
<td>.7003</td>
<td>.7315</td>
</tr>
<tr>
<td>bibtex</td>
<td>.3757</td>
<td>.3838</td>
<td>.3833</td>
<td>.9288</td>
<td>.8840</td>
<td>.7881</td>
</tr>
<tr>
<td>delicious</td>
<td>.1596</td>
<td>.1646</td>
<td>.1661</td>
<td>.9273</td>
<td>.8305</td>
<td>.7901</td>
</tr>
<tr>
<td>bookmarks</td>
<td>.2353</td>
<td>.2435</td>
<td>.2358</td>
<td>.9401</td>
<td>.8842</td>
<td>.7945</td>
</tr>
<tr>
<td>corel5k</td>
<td>.0612</td>
<td>.0584</td>
<td>.0580</td>
<td>.9795</td>
<td>.8503</td>
<td>.7481</td>
</tr>
<tr>
<td>IMDB-F</td>
<td>.1161</td>
<td>.1163</td>
<td>.1160</td>
<td>.8256</td>
<td>.7109</td>
<td>.6411</td>
</tr>
</tbody>
</table>
Future Work

- Experiments with the lowest common subsumer (LCS)
  - Criterion that determines whether or not to reject the LCS
- Exploiting additional resources of the LOD cloud: DBPedia, LinkedGeoData, Geospecies knowledge base, Bio2RDF
- General first check of all labels to detect the domain they are referred to → select the appropriate sense of a label based on this domain
- Discover additional types of relationships among the labels (e.g. mutual exclusion, parent-child relationship among existing labels)
Thank you for your attention!