Cross-training: Learning probabilistic mappings between topics

Sunita Sarawagi  Soumen Chakrabarti  Shantanu Godbole
IIT Bombay

SIGKDD'04

Presented by Gabriel Ripoche
CS591CXZ – Text Mining
October, 13th 2004
Problem – Why cross-training?

Given a set of documents annotated using different taxonomies:

- How can these taxonomies be reconciled?
- How can these taxonomies lead to more accurate classification?

Why is this an interesting problem?

- No global standards
- Evolving taxonomies

Applications:

- Semantic web
- Directories, catalogues, etc.
Concept – What is cross-training?

Cross-training:

1. Use label assignments from taxonomy A to make better inferences about label assignments for taxonomy B

2. Use inferred relationships between A and B to map A and B
Method 1: **EM2D** – Cross-trained naive bayes with EM

Principle:

- Use **expectation maximization** (EM) for semi-supervised learning. (small annotated set + large training set of unlabeled documents)
- Learning two label taxonomies (A, B) = learning over A x B matrix

How to do it?

- In EM update rule, use known label to **restrict contribution** of training document to possible subset of labels
- E.g.: If A_2 is known, only P(B_2 | A_2) should be updated, not all P(B_x | A_y)

Note: Initialization is important! Paper discusses ways to initialize EM
Method 2: SVM-CT – Cross-trained SVM

Principle:

- For each taxonomy, learn SVM classifiers over both document terms + additional features representing labels of taxonomy.
- Iterate + alternate (learn SVM-A over B features, then SVM-B over A features, ...)

How to do it?

- Add $N$ additional features corresponding to each possible label from other taxonomy
- E.g.: When learning SVM for taxonomy B, add $|A|$ features, one for each label in $A = (A_1, A_2, A_3, A_4)$
Training and evaluation settings

- **tune set (small)**: 
  - \((A_1, B_3)\)  
  - \((A_2, B_1)\)  
  - \((A_1, B_3)\)

- **train set (big)**: 
  - \((A_1, ?)\)  
  - \((A_2, ?)\)  
  - \((?, B_3)\)

- **Cross-training**:

- **test set 1**:
  - \((A_1, ?)\)
  - \((A_2, B_3)\)
  - \((A_1, B_3)\)

- **test set 2** (zero mapping):
  - \((?, ?)\)
  - \((? , B_3)\)

- **Output**:
  - \((A_1, B_3)\)
Experiment 1: NB vs. SVM

- Comparing baseline Naive Bayes, SVM, and SVM-CT

- SVM and SVM-CT are about the same
- SVM beats NB in almost all cases
- WHY use NB in rest of experiments if SVM is better?!?
Experiment 2: EM2D vs. A&S

- How does EM2D compare to state-of-the-art?
Experiment 3: EM2D asymmetric

- Can cross-training work if one taxonomy has very few examples?

- Smearing in $D_A - D_B$ over all B labels (don't rely on B label assignments)
- Small damping factor for $D_A - D_B$ (docs in $D_A - D_B$ don't count for much)
- Tuneset used (use reliable data for the real cross-training)
Experiment 4: EM2D zero-label

- Can cross-training guess **both labels** better?

- EM1D = NB + EM (no cross-training)
- EM2D-D = EM2D + model aggregation (?)
- EM2D-G = EM2D + guess (first guess one label, then use it to guess other)
Experiment 5: SVM-CT mapping

- Feature weights in SVM-CT give interesting mappings

1. Learns 1 to 1 mappings
2. Learns 1 to N mappings (parent-child)
3. Learns non existent mappings

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Dmoz.</th>
<th>Maps to Yahoo.</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autos</td>
<td>News&amp;Magazines</td>
<td>News&amp;Media</td>
<td>0.147</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Volkswagen</td>
<td>-0.156</td>
</tr>
<tr>
<td>Movies</td>
<td>Genres/Western</td>
<td>Titles/Western</td>
<td>0.242</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Titles/Horror</td>
<td>-0.052</td>
</tr>
<tr>
<td>Outdoors</td>
<td>Scuba Diving</td>
<td>Scuba</td>
<td>5.878</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Snowmobiling</td>
<td>-0.647</td>
</tr>
<tr>
<td>Photo</td>
<td>Techs&amp;Styles</td>
<td>Pinhole Ph’graphy</td>
<td>2.796</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3D</td>
<td>0.964</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Panoramic</td>
<td>0.921</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Organizations</td>
<td>-1.184</td>
</tr>
<tr>
<td>Software</td>
<td>Accounting</td>
<td>NOTA</td>
<td>0.156</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Screen Savers</td>
<td>0.103</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OS/Unix</td>
<td>-0.171</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Yahoo.</th>
<th>Maps to Dmoz.</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autos</td>
<td>Corvette</td>
<td>Chevrolet</td>
<td>0.981</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Parts&amp;Accessories</td>
<td>-0.266</td>
</tr>
<tr>
<td>Movies</td>
<td>SciFi&amp;Fantasy</td>
<td>Series/Star-Wars</td>
<td>1.123</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reviews</td>
<td>-0.824</td>
</tr>
<tr>
<td>Outdoors</td>
<td>Scuba</td>
<td>Scuba Diving</td>
<td>4.822</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wildlife</td>
<td>-0.437</td>
</tr>
<tr>
<td>Photo</td>
<td>Pinhole Ph’graphy</td>
<td>Techs&amp;Styles</td>
<td>0.4842</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Photographers</td>
<td>-0.270</td>
</tr>
<tr>
<td>Software</td>
<td>OS/MSWindows</td>
<td>OS/MSWindows</td>
<td>0.018</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NOTA</td>
<td>-0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OS/Unix</td>
<td>-0.008</td>
</tr>
</tbody>
</table>
Conclusion, questions, and critiques

Conclusions:

- Method presented for cross-training using multiple taxonomies
- Many experiments cover different situations
- Direct applications in ontology merging, semantic web, ...

Questions & Critiques:

- Compare SVM-CT and EM2D? SVM > NB (1), is SVM-CT > EM2D ?
- Comparing EM2D and A&S (2): A&S is crippled, is that fair?
- Asymmetric scenario (3): isn't their method a hack?
- EM2D-G is best in zero-label (4). Why is that so? Why is it surprising?
- Paper poorly organized