Combining Labeled and Unlabeled Data with Co-Training

Author: Avrim Blum, Tom Mitchell
Report: Weigang Li
Paper’s base information

• *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT-98).*

• ‘<<Machine Learning>>’s author

• Director, [Center for Automated Learning and Discovery School of Computer Science](#) Carnegie Mellon University
Outline

• Co-Training motivation
• What’s Co-Training?
• Co-Training setting
• Input and output of Co-Training
• General process of Co-Training
• Experiments and discussion
• Conclusion
Co-Training motivation

- Most machine learning techniques rely on labeled data
- But labeled data is expensive
- Unlabeled data is plentiful
- How to boost performance of a learning algorithm when only a small set of labeled data?
- Co-Training is the one of these algorithms
Outline

• Co-Training motivation
• What’s Co-Training?
• Co-Training setting
• Input and output of Co-Training
• General process of Co-Training
• Experiments and discussion
• Conclusion
What’s co-training

• Co-training is a weakly supervised learning paradigm in which the redundancy of the learning task is captured by training two classifiers using separate views of the same data.
Outline

• Co-Training motivation
• What’s Co-Training?
• Co-Training setting
• Input and output of Co-Training
• General process of Co-Training
• Experiments and discussion
• Conclusion
Co-Training setting (Where to use it)

• Dataset has a natural division of its features

• Two assumptions
  – The instances distribution is compatible with the target function
    • two classifiers label one document into same class
  – The features in one set of an instance are conditional independent of the features in the second set
    • As informative as a random document
A formal framework

• If problem setting provides redundantly sufficient features, classifier are conditional independence

\[ \text{learn } f : X \rightarrow Y \]
\[ \text{where } X = X_1 \times X_2 \]
\[ \text{where } x \text{ drawn from unknown distribution} \]
\[ \exists f_1, f_2 \quad (\forall x) f_1(x_1) = f_2(x_2) = f(x) \]
One practical application

• Web-page classification is an example
• CS faculty member pages or course home pages at University
• An interesting feature:
  – The text appearing on the document itself
  – The anchor text attached to hyperlinks pointing to this page
Outline

- Co-Training motivation
- What’s Co-Training?
- Co-Training setting
- Input and output of Co-Training
- General process of Co-Training
- Experiments and discussion
- Conclusion
Input and output of co-training

• Input:
  – labeled data L (a small set of labeled web pages)
  – unlabeled data U (large set of unlabeled web pages)

• Output:
  – Label the unlabeled data (classify the unlabeled documents)
Outline

• Co-Training motivation
• What’s Co-Training?
• Co-Training setting
• Input and output of Co-Training
• **General process of Co-Training**
• Experiments and discussion
• Conclusion
Underlying classifier of NBC

• Naïve Bayes Classifier, can attain:
  – The posteriori probabilities
    \[ P(w_t|c_j) = \frac{1 + \sum_{i=1}^{|D|} N(w_t, d_i)P(c_j|d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N(w_s, d_i)P(c_j|d_i)}, \]
  – The prior probabilities
    \[ P(c_j) = \frac{1 + \sum_{i=1}^{|D|} P(c_j|d_i)}{|C| + |D|}. \]

• Output:
    \[ P(c_j|d_i) \propto P(c_j)P(d_i|c_j) \]
    \[ = P(c_j) \prod_{k=1}^{|d_i|} P(w_{d_i,k}|c_j). \]
Co-Training Algorithm

• Given
  – labeled data L,
  – unlabeled data U

• Create a pool U’ of examples at random from U

• Loop for \( k \) iterations:
  – Train \( f_1 \) (hyperlink classifier) using L
  – Train \( f_2 \) (page classifier) using L
  – Allow \( f_1 \) to label \( p \) positive, \( n \) negative examples from U’
  – Allow \( f_2 \) to label \( p \) positive, \( n \) negative examples from U’
  – Add these self-labeled examples to L
  – Randomly choose \( 2p+2n \) examples from U to replenish U’
Outline

• Co-Training motivation
• What’s Co-Training?
• Co-Training setting
• Input and output of Co-Training
• General process of Co-Training
• Experiments and discussion
• Conclusion
Comparison

• Co-Training
  – Begin with 12 labeled web pages (academic course)
  – $P=1$, $n=3$, $k=30$, $u=75$

• Supervised Naïve Bayes classifiers
  – Begin with 12 labeled web pages, too

• Three classifiers
  – Hyperlink-based classifier
  – Page-based classifier
  – Combined classifier (multiplying the probabilities)
Co-Training: experiment data

- 1051 web pages from CS at four university
- Hand labeling these pages
- Task:
  - Categories “course home page” as the target function, 22% of the them were course pages
  - 3 positive, 9 negative as L
  - 263 of the 1051 were as a test set
  - Others are unlabeled data
Experimental results

<table>
<thead>
<tr>
<th></th>
<th>Page-based classifier</th>
<th>Hyperlink-based classifier</th>
<th>Combined classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised training</td>
<td>12.9</td>
<td>12.4</td>
<td>11.1</td>
</tr>
<tr>
<td>Co-training</td>
<td>6.2</td>
<td>11.6</td>
<td>5.0</td>
</tr>
</tbody>
</table>

- average error: learning from labeled data 11.1%
- average error: co-training 5.0%
- Page-based is helpful by Co-Training
- Hyperlink-based classifier is helpless by Co-Training
  - The fact that hyperlinks contain fewer words and less capable of expression
Explanation

• Theoretical proof in the paper
  – PAC-learning (probably approximate correct)

• Intuition explanation
  – One classifier finds an “easily classified” pages which maybe difficult for the another classifier
  – Provide useful information each other
Explanation (cont.)

• Supervised NBC
  – Not using the unlabeled data information
  – Directly using the probabilities

• Co-Training
  – Using split features
  – Ranking the documents by confidence
  – Incrementally using the unlabeled data
Some questions

• The model is an over-simplification of real-world target functions and distributions
• Conditional independence is a somewhat unreasonably strict assumption
• Experiment involves just one data set and one target function
Outline

• Co-Training motivation
• What’s Co-Training?
• Co-Training setting
• Input and output of Co-Training
• General process of Co-Training
• Experiments and discussion
• Conclusion
Conclusions

• Unlabeled data improves supervised learning when example features are redundantly sufficient

• Some Theoretical results
Other applications

• IE (Riloff and Jones, 1999)
  – A term matching classifier over word tokens
  – A context rule classifier over the neighboring words of the tokens

• WSD (Yarowsky, 1995)
  – A sense classifier using the local context of the word
  – A classifier based on the sense of other occurrences of that word in the same document

• NER (Collins & Singer, 1999)
  – The spelling of the named entity
  – The context in which the entity occurs

• Parsing
  – ........
Thanks
Any questions?