Semi-Supervised Learning for Document Classification

Anastasia Krithara

Xerox Research Centre Europe
LIP6 - Pierre and Marie Curie University (Paris VI)

MLSS 2007
Motivation

Supervised Learning:

Given a training set \( \{(x_i, y_i)\} \), estimate a decision function (a probability \( P(y|x) \))

Problem:
- The annotation process is often costly and time-consuming...

\[\xrightarrow{\text{Semi-Supervised Learning}}\]
Supervised Learning:

Given a training set \( \{(x_i, y_i)\} \), estimate a decision function (a probability \( P(y|x) \))

Problem:

- The annotation process is often costly and time-consuming...

\[\rightarrow\] Semi-Supervised Learning
Motivation

**Supervised Learning:**
Given a training set \( \{(x_i, y_i)\} \), estimate a decision function (a probability \( P(y|x) \)).

**Problem:**
- The annotation process is often costly and time-consuming...

\[ \rightarrow \textbf{Semi-Supervised Learning} \]
Semi-Supervised Learning (SSL)
- semi-supervised PLSA (ssPLSA)
  - ssPLSA with a “Fake label” model
  - ssPLSA with a mislabeling error model

Evaluation
- Experiments
- Results

Conclusion
Outline

1 Semi-Supervised Learning (SSL)
   - semi-supervised PLSA (ssPLSA)
     - ssPLSA with a “Fake label” model
     - ssPLSA with a mislabeling error model

2 Evaluation
   - Experiments
   - Results

3 Conclusion
Semi-Supervised Learning (SSL)

Supervised Learning:
Given a training set \( \{(x_i, y_i)\} \), estimate a decision function (a probability \( P(y|x) \)).

Semi-Supervised Learning (SSL):
Same goal as in supervised learning but in addition, a set of unlabeled data \( x_i \) is available (in general unlabeled data \( \gg \) labeled data).

Unlabeled data can give us some valuable information about \( P(X) \).
Supervised Learning:
Given a training set \( \{(x_i, y_i)\} \), estimate a decision function (a probability \( P(y|x) \))

Semi-Supervised Learning (SSL):
Same goal as in supervised learning but in addition, a set of unlabeled data \( x_i \) is available (in general unlabeled data \( \gg \) labeled data)

Unlabeled data can give us some valuable information about \( P(X) \)
Semi-Supervised Learning (SSL):

Supervised Learning:
Given a training set \{ (x_i, y_i) \}, estimate a decision function (a probability \( P(y|x) \)).

Semi-Supervised Learning (SSL):
Same goal as in supervised learning but in addition, a set of unlabeled data \( x_i \) is available (in general unlabeled data >> labeled data).

Unlabeled data can give us some valuable information about \( P(X) \).
We represent our document collection as co-occurrences of documents and terms.
Problems

- Synonyms: different words have the same meaning
- Polysems: words with multiple meanings
  \[\Rightarrow\] Disconnection between topics and words

Solution

PLSA aims to discover something about the meaning behind the words, about the topics of the document.
Probabilistic Latent Semantic Analysis (PLSA)

Problems

- Synonyms: different words have the same meaning
- Polysems: words with multiple meanings
  \[\Rightarrow\] Disconnection between topics and words

Solution

PLSA aims to discover something about the \textit{meaning} behind the words, about the \textit{topics} of the document.
We model our data using a mixture model, under the assumption that \( d \) and \( w \) are independent:

\[
P(w, d) = P(d) \sum_{\alpha} P(w|\alpha)P(\alpha|d)
\]

\((\alpha = 1 \ldots A \text{ is the index over } A \text{ latent components})\)

- \( P(w|\alpha) \Rightarrow \text{the profile of a topic (component)} \)
- \( P(\alpha|d) \Rightarrow \text{the topics of a document} \)
We model our data using a mixture model, under the assumption that $d$ and $w$ are independent:

$$P(w, d) = P(d) \sum_{\alpha} P(w|\alpha)P(\alpha|d)$$

($\alpha = 1 \ldots A$ is the index over $A$ latent components)

- $P(w|\alpha) \Rightarrow$ the profile of a topic (component)
- $P(\alpha|d) \Rightarrow$ the topics of a document
**ssPLSA with a "fake label" model**

When the ratio of labeled and unlabeled documents is very low:

\[ \rightarrow \] Some components may contain only unlabeled examples

- In this case, arbitrary probabilities will be assigned to these components

**Solution**

Introduce an additional "fake" label \( Z_0 \)

- All labeled examples keep their own label
- All unlabeled examples get the new "fake" label
Semi-Supervised Learning (SSL)  
Semi-Supervised PLSA  

**ssPLSA with a "fake label" model**

**Model**
- Parameters:
  \[ \Lambda = \{ p(\alpha | d), p(y | \alpha), p(w | \alpha) : \alpha \in A, d \in D, w \in W \} \]
- Log-likelihood:
  \[ L_1 = \sum_{x \in \mathcal{Z}_l \cup \mathcal{X}_u} \log p(x, y) = \sum_{x \in \mathcal{Z}_l \cup \mathcal{X}_u} \log p(w, d, y) \]
- EM (Expectation-Maximization) algorithm

**"Fake labels"**

We distribute the probability obtained for the "fake label" on the "true" ones:

\[ P(y|x) \propto \sum_{\alpha} P(\alpha|x)P(y|\alpha) + \lambda \sum_{\alpha} P(\alpha|x)P(y=0|\alpha) \]

where \( \lambda << 1 \) and \( y = 1, \cdots, K \)
**ssPLSA with a mislabeling error model**

- For all unlabeled data we assume that there exists:
  - a perfect label (the true one $y$)
  - an imperfect label (the estimated one $\tilde{y}$)
- We model these labels by the following probabilities:
  $$\forall (k, h) \in C \times C, \beta_{kh} = p(\tilde{y} = k | y = h)$$
  subject to the constraint that
  $$\forall h, \sum_k \beta_{kh} = 1$$

![Diagram showing labeled and unlabeled documents](image)
ssPLSA with a mislabeling error model

Model

- **Parameters:**
  \[ \Phi = \{ p(\alpha | d), p(w | \alpha), \beta_{\tilde{y}|y} : d \in D, w \in W, \alpha \in A, y \in C, \tilde{y} \in C \} \]

- **Log-likelihood:**
  \[ L_2 = \sum_{d \in D_l} \sum_{w} n(w, d) \log \sum_{\alpha} p(d)p(w|\alpha)p(\alpha|d)p(y|\alpha) + \sum_{d \in D_u} \sum_{w} n(w, d) \log p(w, d, \tilde{y}) \]

- **EM algorithm**
Outline

1. Semi-Supervised Learning (SSL)
   - semi-supervised PLSA (ssPLSA)
     - ssPLSA with a “Fake label” model
     - ssPLSA with a mislabeling error model

2. Evaluation
   - Experiments
   - Results

3. Conclusion
Experiments

Characteristics of the datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>20Newsgroups</th>
<th>WebKB</th>
<th>Reuters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size of the collection</td>
<td>20000</td>
<td>4196</td>
<td>4381</td>
</tr>
<tr>
<td># of classes, $K$</td>
<td>20</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Size of the vocabulary, $</td>
<td>W</td>
<td>$</td>
<td>38300</td>
</tr>
<tr>
<td>Training set, $</td>
<td>D_l \cup D_u</td>
<td>$</td>
<td>16000</td>
</tr>
<tr>
<td>Test set</td>
<td>4000</td>
<td>839</td>
<td>876</td>
</tr>
</tbody>
</table>

Evaluation measures

We calculate the F-score: $F = \frac{2PR}{P+R}$

$P \Rightarrow$ Precision (ratio of true positives over all returns)

$R \Rightarrow$ Recall (ratio of true positives over all positives)
Results

Evaluation Results

F-Score (y-axis) versus, the percentage of labeled examples in the training set | graphs for the various algorithms on 20Newsgroups.

20 Newsgroups

F-Score

% of the labeled data in the training set

20 Newsgroups

F-Score

% of the labeled data in the training set

24 August 2007
Results
Outline

1 Semi-Supervised Learning (SSL)
   - semi-supervised PLSA (ssPLSA)
     - ssPLSA with a “Fake label” model
     - ssPLSA with a mislabeling error model

2 Evaluation
   - Experiments
   - Results

3 Conclusion
Motivation

- Reduce the annotation cost for the text classification task

Work presented

- Two semi-supervised variants of the PLSA algorithm
  - ssPLSA with a “fake label” model
  - ssPLSA with a mislabeling error model
- Evaluation of the above algorithms
Thank you

Questions?