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...

$\Longrightarrow$ ***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...

$\Longrightarrow$ ***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...

$\Longrightarrow$ ***Semi-Supervised Learning***

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

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 $>>$ 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 $P(X)$

Semi-Supervised Learning (SSL)

Supervised Learning:

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

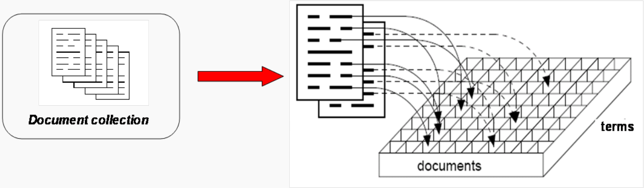
Semi-Supervised Learning (SSL):

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 $P(X)$

Data representation

We represent our document collection as co-occurences of documents and terms



Probabilistic Latent Semantic Analysis (PLSA)

Problems

- Synonyms: different words have the same meaning
- Polysems: words with multiple meanings
  $\Longrightarrow$ 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
  $\Longrightarrow$ 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)

- We model our data using a mixture model, under the assumption that d and w are independent:
  $P(w,d) = P(d)\sum_{\alpha} P(w|\alpha)P(a|d)$
  ($\alpha = 1 \dots A$ 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

Probabilistic Latent Semantic Analysis (PLSA)

- We model our data using a mixture model, under the assumption that d and w are independent:
  $P(w, d) = P(d) \sum_{\alpha} P(w|\alpha)P(a|d)$
  ($\alpha = 1 \ldots A$ 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

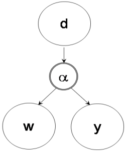
When the ratio of labeled and unlabeled documents is very low:
$\Longrightarrow$ Some components may contain only unlabeled examples

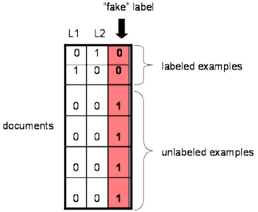
- In this case, arbitrary probabilities will be assigned to these components

Solution

Introduce an additional "fake" label Z_0

- All labeled examples keep their own label
- All unlabeled examples get the new "fake" label





ssPLSA with a "fake label" model

Model

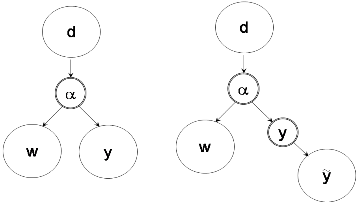
- Parameters:
  $\Lambda = \{p(\alpha \mid d), p(y \mid \alpha), p(w \mid \alpha) : \alpha \in A, d \in \mathcal{D}, w \in \mathcal{W}\}$
- Log-likelihood:
  $\mathcal{L}_1 = \sum_{x \in \mathcal{Z}_l \cup \mathcal{X}_u} \log p(x, y) = \sum_{x \in \mathcal{Z}_l \cup \mathcal{X}_u} \log p(w, d, y)$
- 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:
  $\Rightarrow$ a perfect label (the true one y)
  $\Rightarrow$ an imperfect label (the estimated one $\tilde{y}$)
- We model these labels by the following probabilities:
  $\forall (k,h) \in \mathcal{C} \times \mathcal{C}, \beta_{kh} = p(\tilde{y} = k | y = h)$ subject to the constaint that $\forall h, \sum_k \beta_{kh} = 1$



Labeled documents Unlabeled documents

ssPLSA with a mislabeling error model

Model

- Parameters:
  $\Phi = \{p(\alpha \mid d), p(w \mid \alpha), \beta_{\tilde{y}|y} : d \in \mathcal{D}, w \in \mathcal{W}, \alpha \in A, y \in \mathcal{C}, \tilde{y} \in \mathcal{C}\}$
- Log-likelihood:
  $\mathcal{L}_2 = \sum_{d \in D_l} \sum_w n(w,d) \log \sum_\alpha p(d)p(w|\alpha)p(\alpha|d)p(y|\alpha)$
  $+ \sum_{d \in D_u} \sum_w n(w,d) \log p(w,d,\tilde{y})$
- EM algorithm

Outline

Experiments

Characteristics of the datasets

Dataset	20Newsgroups	WebKB	Reuters
Size of the collecion	20000	4196	4381
# of classes, K	20	4	7
Size of the vocabulary, $\lvert\mathcal{W}\rvert$	38300	9400	4749
Training set, $\lvert D_l \cup D_u\rvert$	16000	3257	3504
Test set	4000	839	876

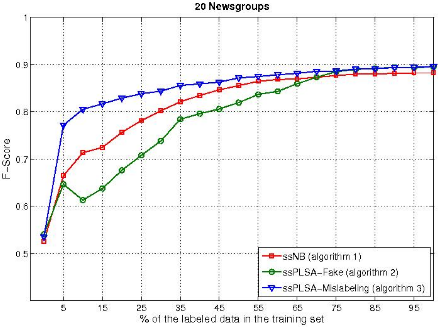
Evaluation measures

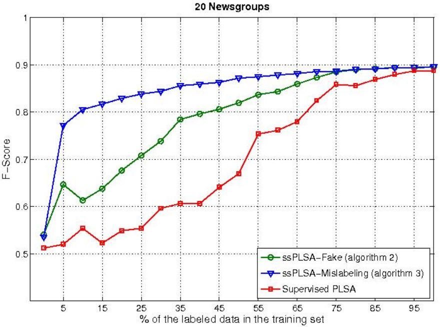
We calculate the F-score: $F = \frac{2PR}{P+R}$

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

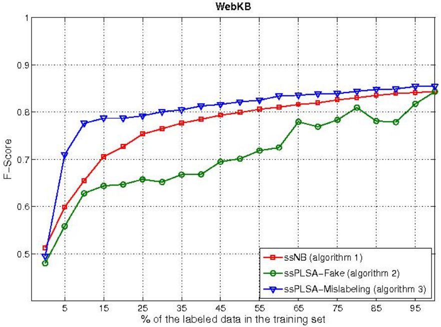
$R \Rightarrow$ Recall (ratio of true positives over all positives)

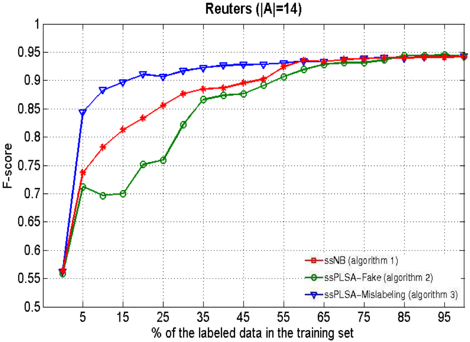
Results


20 Newsgroups
F−Score
% of the labeled data in the training set
ssNB (algorithm 1)
ssPLSA−Fake (algorithm 2)
ssPLSA−Mislabeling (algorithm 3)


20 Newsgroups
F−Score
% of the labeled data in the training set
ssPLSA−Fake (algorithm 2)
ssPLSA−Mislabeling (algorithm 3)
Supervised PLSA

Results


WebKB
F−Score
% of the labeled data in the training set
ssNB (algorithm 1)
ssPLSA−Fake (algorithm 2)
ssPLSA−Mislabeling (algorithm 3)


Reuters (|A|=14)
F-score
% of the labeled data in the training set
ssNB (algorithm 1)
ssPLSA−Fake (algorithm 2)
ssPLSA−Mislabeling (algorithm 3)

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

Summary

Motivation

- Reduce the annotation cost for the text classicafication 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?