Semi-supervised learning

LING 572

Fei Xia

Week 10: 3/11/2010

Outline

- Overview of Semi-supervised learning (SSL)

- Self-training

Additional Reference

- Xiaojin Zhu (2006): Semi-supervised learning literature survey.

- Olivier Chapelle et al. (2005): Semi-supervised Learning. The MIT Press.

Overview of SSL

What is SSL?

- Labeled data:
  - Ex: POS tagging: tagged sentences
  - Creating labeled data is difficult, expensive, and/or time-consuming.
- Unlabeled data:
  - Ex: POS tagging: untagged sentences.
  - Obtaining unlabeled data is easier.
- Goal: use both labeled and unlabeled data to improve the performance

- Learning
  - Supervised (labeled data only)
  - Semi-supervised (both labeled and unlabeled data)
  - Unsupervised (unlabeled data only)

- Problems:
  - Classification
  - Regression
  - Clustering
  - …

➔ Focus on semi-supervised classification problem

A brief history of SSL

- The idea of self-training appeared in the 1960s.
- SSL took off in the 1970s.
- The interest for SSL increased in the 1990s, mostly due to applications in NLP.

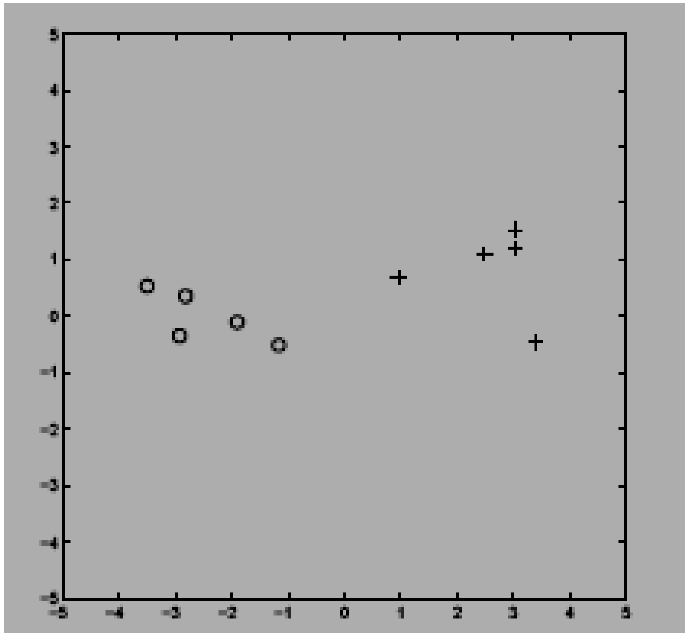
Does SSL work?

- Yes, under certain conditions.
  - The problem itself: the knowledge on p(x) carry information that is useful for the inference of p(y | x).
  - Algorithm: the modeling assumption fits well with the problem structure.

- SSL will be most useful when there are far more unlabeled data than labeled data.

- SSL could degrade the performance when mistakes reinforce themselves.

Illustration
(Zhu, 2006)



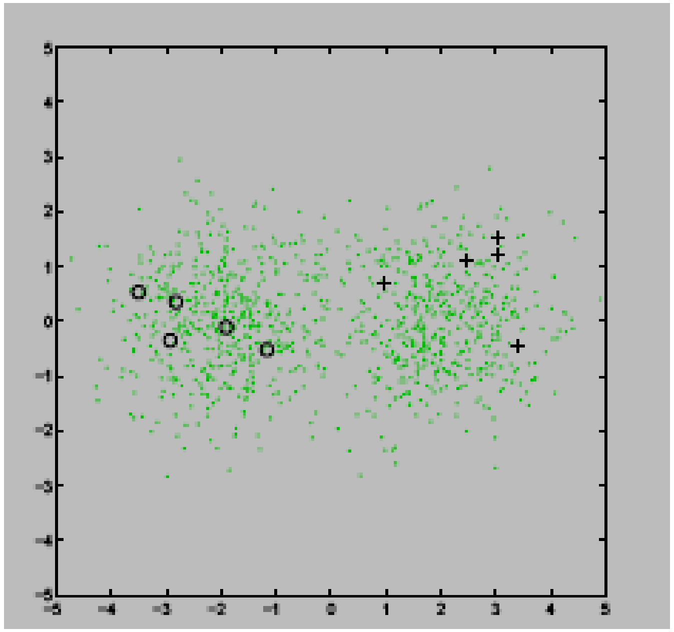
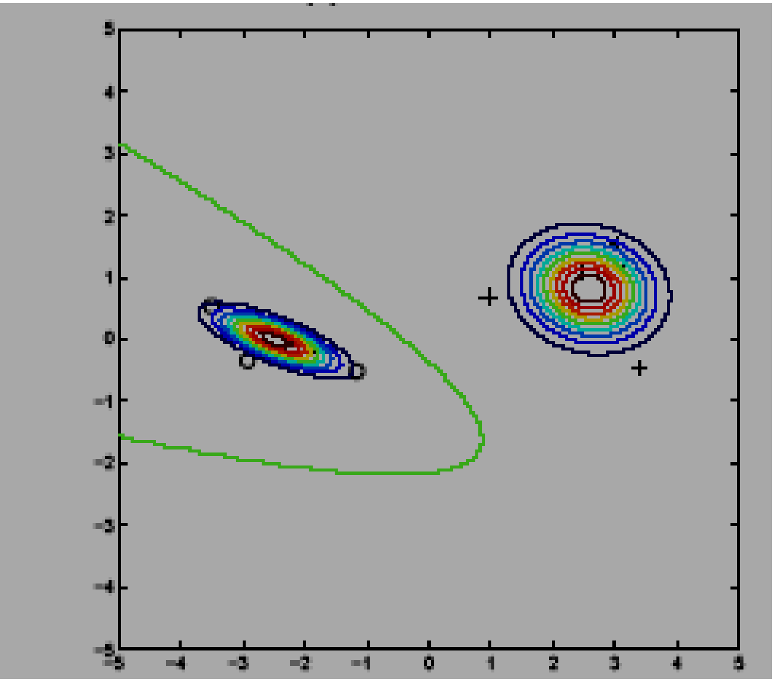
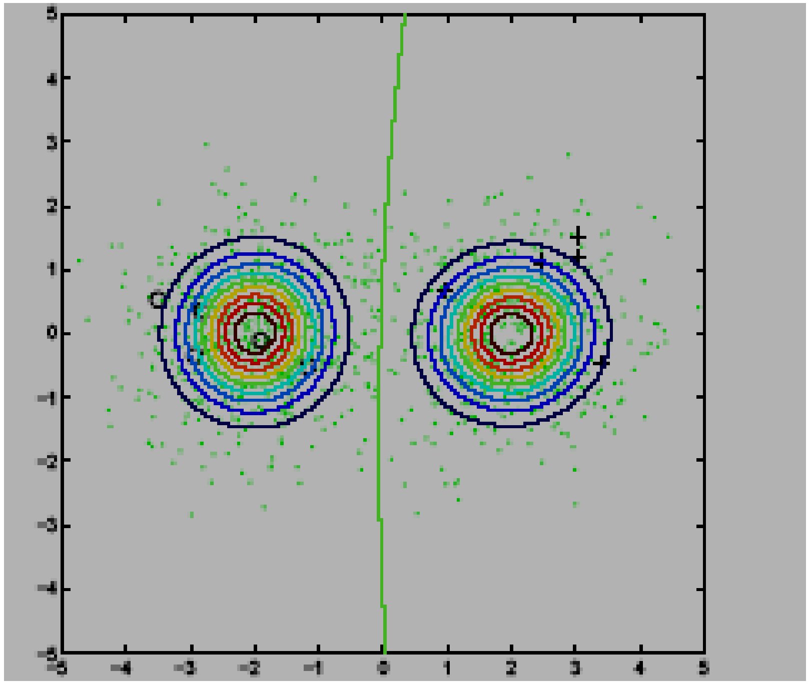


Illustration (cont)





Assumptions

- Smoothness (continuity) assumption: if two points x_1 and x_2 **in a high-density region** are close, then so should be the corresponding outputs y_1 and y_2.

- Cluster assumption: If points are in the same cluster, they are likely to be of the same class.

  Low density separation: the decision boundary should lie in a low density region.

- ….

SSL algorithms

- Self-training
- Co-training
- Generative models:
  - Ex: EM with generative mixture models
- Low Density Separations:
  - Ex: Transductive SVM
- Graph-based models

Which SSL method should we use?

- It depends.

- Semi-supervised methods make strong model assumptions.

- Choose the ones whose assumptions fit the problem structure.

Self-training

Basics of self-training

- Probably the earliest SSL idea.

- Also called self-teaching or bootstrapping.

- Appeared in the 1960s and 1970s.

- First well-known NLP paper: (Yarowsky, 1995)

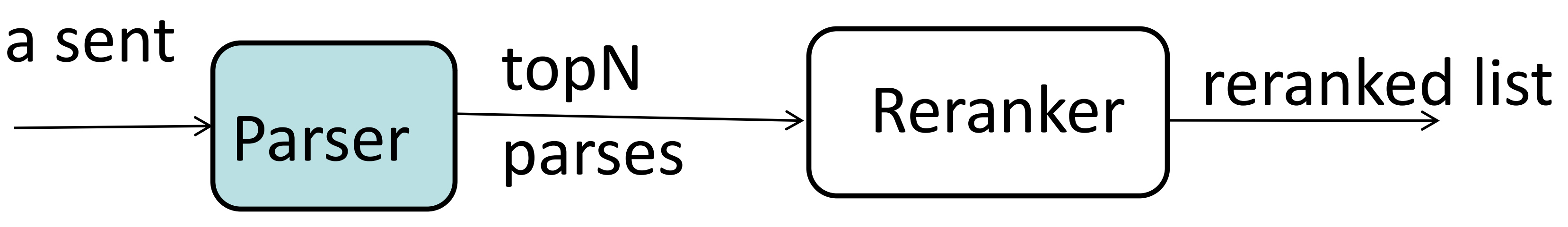
Self-training algorithm

- Let L be the set of labeled data, U be the set of unlabeled data.

- Repeat
  - Train a classifier h with training data L
  - Classify data in U with h
  - Find a subset U' of U with the most confident scores.
  - L + U' ➔ L
  - U – U' ➔ U

An example: (McClosky et al., 2006)

- Setting:
  - Training data:
    - Labeled data: WSJ
    - Unlabeled data: NANC
  - Test data: WSJ



The procedure

- Self-training procedure:
  - Train a stage-1 parser and a reranker with WSJ data
  - Parse NANC data and add the best parse to re-train stage-1 parser

- Best parses for NANC sentences come from
  - the stage-1 parser (“Parser-best”)
  - the reranker (“Reranker-best”)

Sentences added	Parser-best	Reranker-best
0 (baseline)	90.3	
50k	90.1	90.7
250k	90.1	90.7
500k	90.0	90.9
750k	89.9	91.0
1,000k	90.0	90.8
1,500k	90.0	90.8
2,000k	–	91.0

Conclusion:

- Self-training alone does not help
- Self-training with reranking provides a modest gain

Summary of self-training

- The algorithm is straightforward and intuitive.

- It could produce good results.

- Added unlabeled data pollute the original labeled data

Papers on self-training

- Yarowsky (1995): WSD
- Riloff et al. (2003): identify subjective nouns
- Maeireizo et al. (2004): classify dialogues as “emotional” or “non-emotional”.
- McClosky et al. (2006): combine self-training and reranking for parsing

Summary

- SSL uses both labeled and unlabeled data.
- There are many SSL algorithms.
- SSL algorithms can improve the performance if the data satisfies the assumption made by the algorithms.
- An example: self-training

Additional slides

Semi-supervised and active learning

- They address the same issue: labeled data are hard to get.
- Semi-supervised: choose the unlabeled data to be added to the labeled data.
- Active learning: choose the unlabeled data to be annotated.

Self-training for parsing adaptation

Sentences added	Parser	Reranking Parser
Baseline BROWN	86.4	87.4
Baseline WSJ	83.9	85.8
WSJ+50k	84.8	86.6
WSJ+250k	85.7	87.2
WSJ+500k	86.0	87.3
WSJ+750k	86.1	87.5
WSJ+1,000k	86.2	87.3
WSJ+1,500k	86.2	87.6
WSJ+2,000k	86.1	87.7
WSJ+2,500k	86.4	87.7

➔ Adding NANC data helps: 83.9% => 86.4%

Co-training

Basic ideas

- The original paper: (Blum and Mitchell, 1998)
- Two “independent” views: split the features into two sets.
  - The instance space: $X = X_1 \times X_2$
  - Each example: $x = (x_1, x_2)$
- Train a classifier on each view.
- Data classified by one classifier can be used to train the other classifier and vice versa.

An example

- Web-page classification: e.g., find homepages of faculty members.
  - Page text: words occurring on that page

    e.g., “research interest”, “teaching”

  - Hyperlink text: words occurring in hyperlinks that point to that page:

    e.g., “my advisor”

Co-training algorithm

Given:

- a set L of labeled training examples
- a set U of unlabeled examples

Create a pool U' of examples by choosing u examples at random from U

Loop for k iterations:

Use L to train a classifier h_1 that considers only the x_1 portion of x

Use L to train a classifier h_2 that considers only the x_2 portion of x

Allow h_1 to label p positive and n negative examples from U'

Allow h_2 to label p positive and n negative examples from U'

Add these self-labeled examples to L

Randomly choose $2p + 2n$ examples from U to replenish U'