Semisupervised learning

learning from labeled and unlabeled data

Saroj K. Meher

Systems Science and Informatics Unit,

Indian Statistical Institute, Bangalore

Outlines

- Why semisupervised learning (SSL)?
- What is SSL?
- How SSL works?
  - Self-training
  - Co-training
- Pros and cons
- Some approaches
- Results and discussion
- Conclusions

Why Semi-Supervised Learning (SSL)?

- Labeled data: labeling usually
  - . . . requires experts
  - . . . costs time
  - . . . is boring
  - . . . requires measurements and devices
  - . . . costs money

  ➔ scarce, expensive

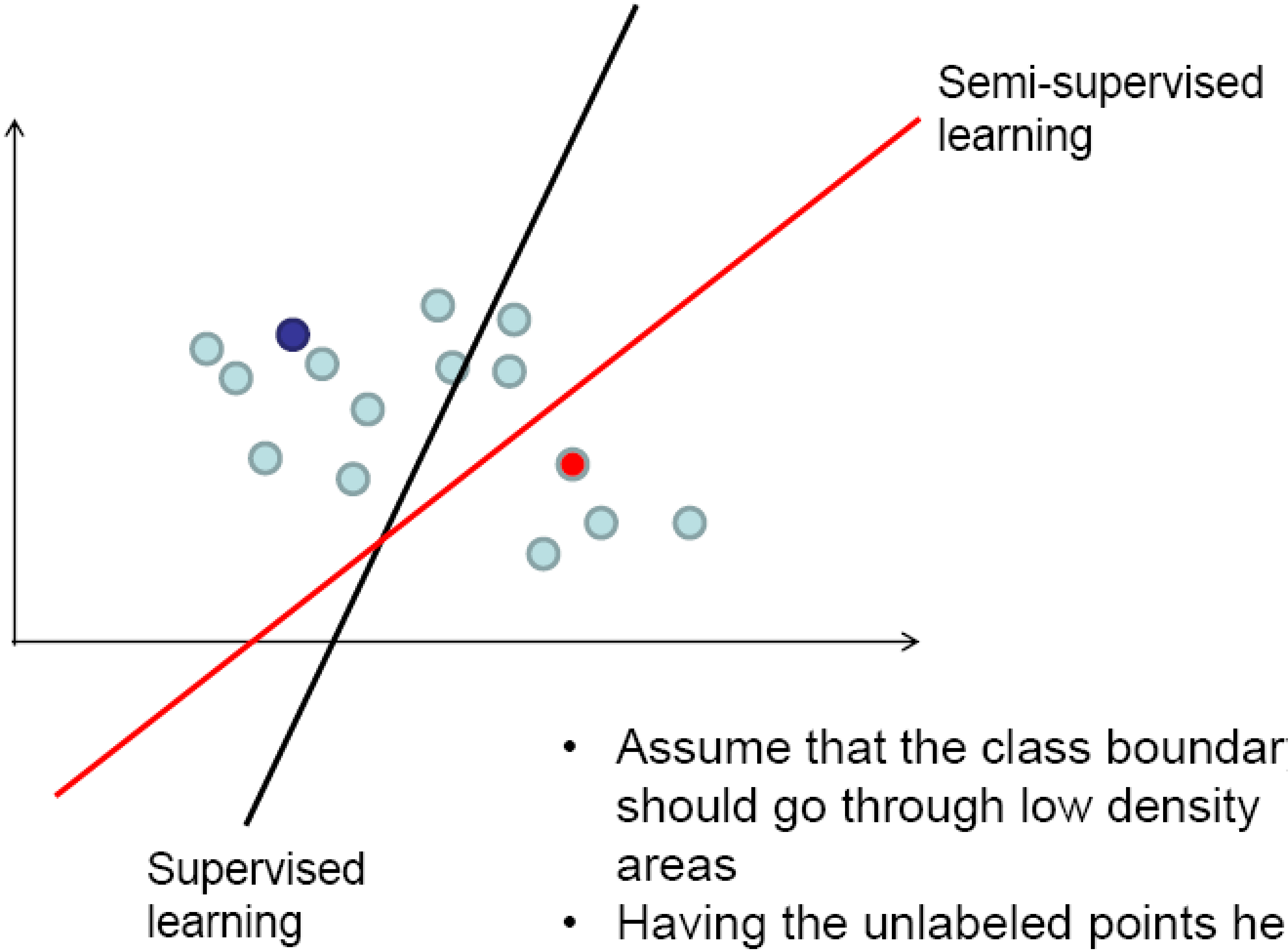
- Unlabeled data: can often be
  - . . . measured automatically
  - . . . found on the web
  - . . . retrieved from databases and collections

  ➔abundant, cheap . . . “for free”

Why Semi-Supervised Learning (SSL)?

- Unsupervised and Supervised learning
  - Two extreme learning paradigms
  - Unsupervised learning
    - e.g., collection of documents without any labels
    - easy to collect
  - Supervised learning
    - each object labeled with a class.
    - expensive to do
- Real life applications are somewhere in between
  – Semi-supervised Learning

Why can unlabeled data help?



- Assume that the class boundary should go through low density areas
- Having the unlabeled points help identify better decision boundary

Examples

Web page / image classification

labeled:

- someone has to read the text
- labels may come from huge ontologyies
- hence has to be done faithfully

unlabeled:

- billions available at no cost

Protein function prediction from sequence

labeled:

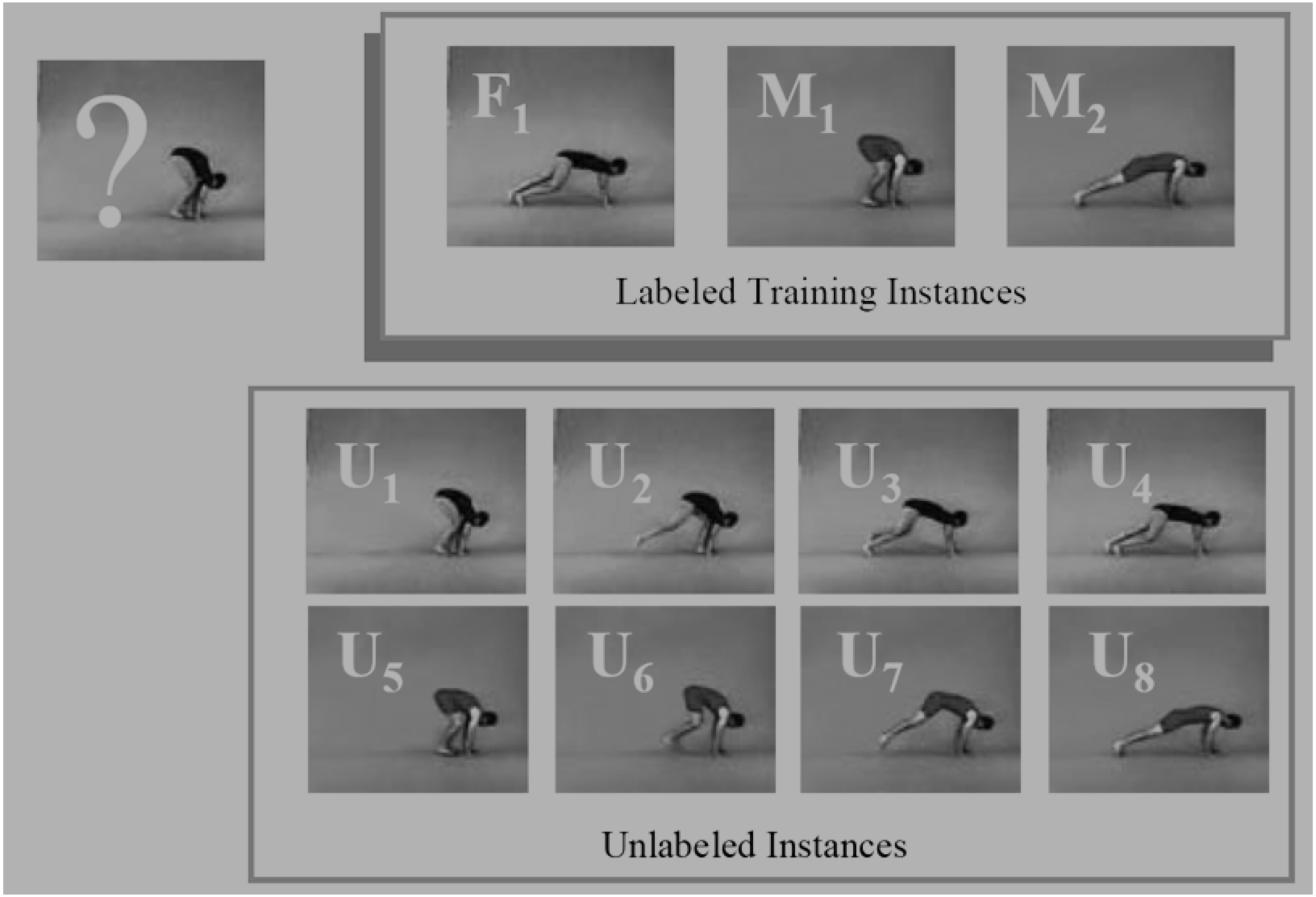
- measurement requires human ingenuity
- can take years for a single label!

unlabeled:

- protein sequences can be predicted from DNA
- DNA sequencing now industrialized

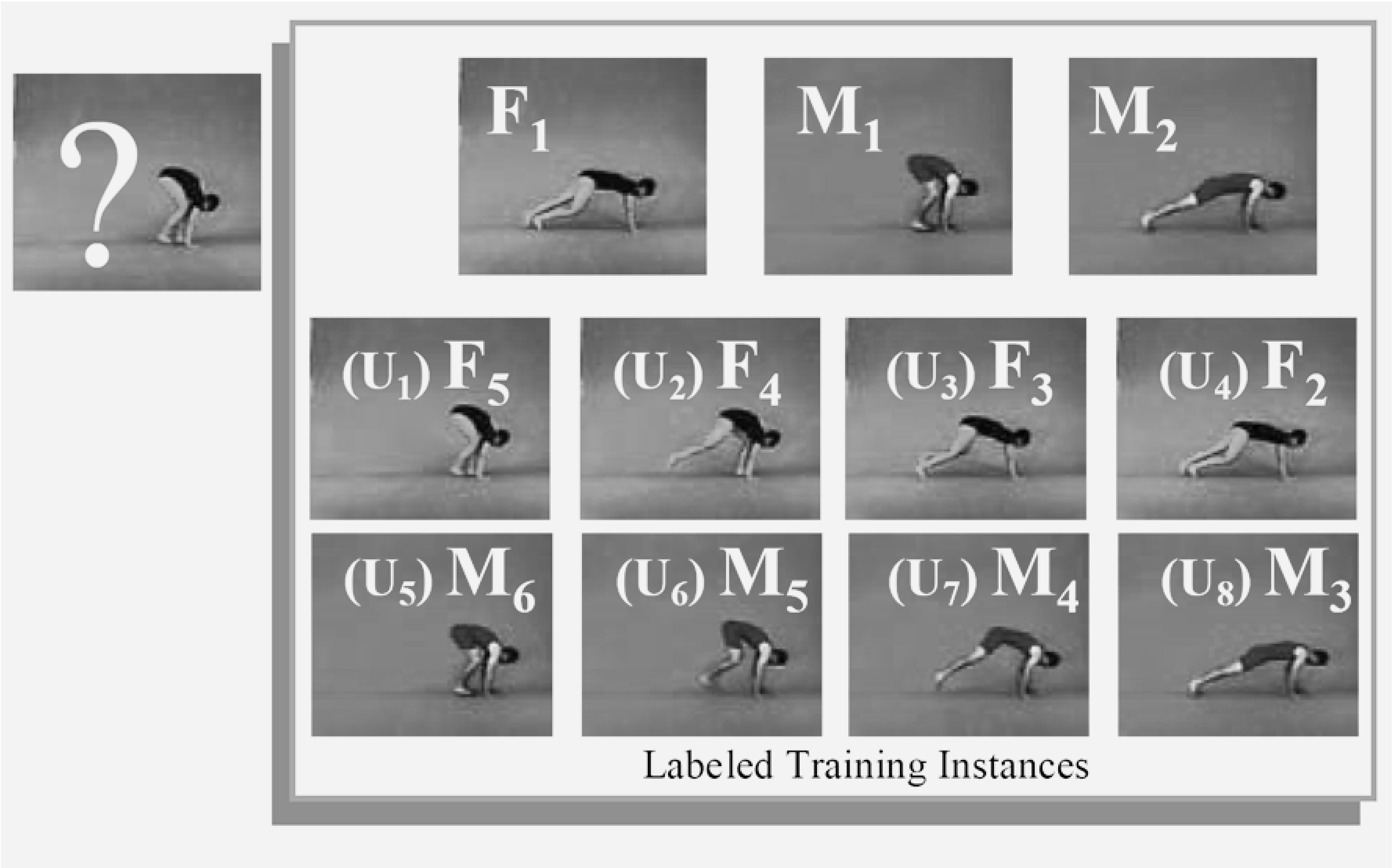
➔ millions available

Example: in action !



Ref: Li Wei and Eamonn Keogh (2006) Semi-Supervised Time Series Classification. SIGKDD 2006.

Example: in action !



Ref: Li Wei and Eamonn Keogh (2006) Semi-Supervised Time Series Classification. SIGKDD 2006.

Semisupervised Learning

- Overview of clustering and classification
- What is semi-supervised learning?
  - Semi-supervised clustering
  - Semi-supervised classification

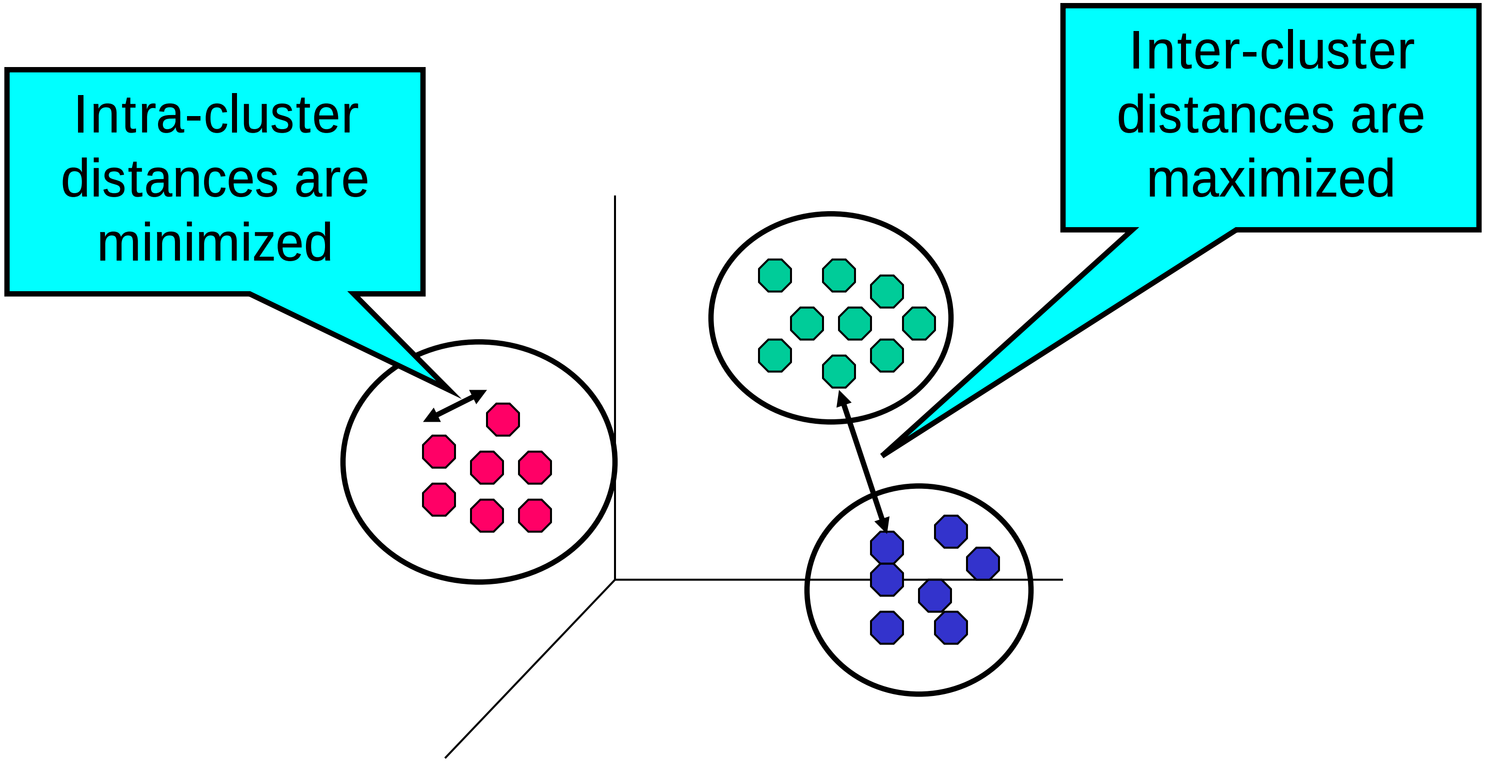
Supervised classification versus unsupervised clustering

- Unsupervised clustering Group similar objects together to find clusters
  - Minimize intra-class distance
  - Maximize inter-class distance

- Supervised classification Class label for each training sample is given
  - Build a model from the training data
  - Predict class label on unseen future data points

What is clustering?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



What is Classification?

Tid	Attrib1	Attrib2	A	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Learning algorithm

Model

Training Set

Test Set

Semi-Supervised Learning

- Combines labeled and unlabeled data during training to improve performance:
  - Semi-supervised classification: Training on labeled data exploits additional unlabeled data, frequently resulting in a more accurate classifier.
  - Semi-supervised clustering: Uses small amount of labeled data to aid and bias the clustering of unlabeled data.

Unsupervised clustering ↔ Semi-supervised learning ↔ Supervised classification

Semi-Supervised Classification

- Algorithms:
  - Semisupervised EM [Ghahramani:NIPS94,Nigam:ML00].
  - Co-training [Blum:COLT98].
  - Transductive SVM’s [Vapnik:98,Joachims:ICML99].
  - Graph based algorithms
- Assumptions:
  - Known, fixed set of categories given in the labeled data.
  - Goal is to improve classification of examples into these known categories.

SSL clustering: problem definition

- **Input:**
  - A set of unlabeled objects, each described by a set of attributes (numeric and/or categorical)
  - A small amount of domain knowledge
- **Output:**
  - A partitioning of the objects into k clusters (possibly with some discarded as outliers)
- **Objective:**
  - Maximum intra-cluster similarity
  - Minimum inter-cluster similarity
  - High consistency between the partitioning and the domain knowledge

Why semi-supervised clustering?

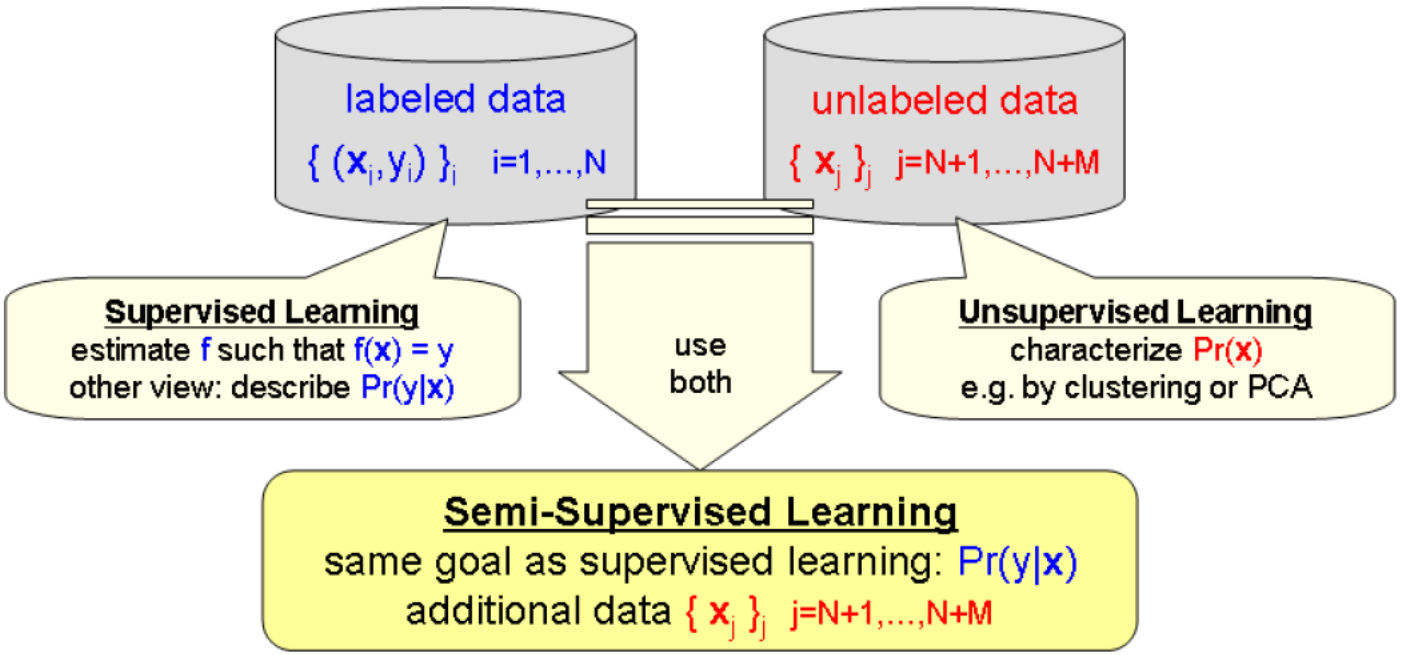
- **Why not clustering?**
  - The clusters produced may not be the ones required.
  - Sometimes there are multiple possible groupings.
- **Why not classification?**
  - Sometimes there are insufficient labeled data.
- **Potential applications**
  - Bioinformatics (gene and protein clustering)
  - Document hierarchy construction
  - News/email categorization
  - Image categorization

Semi-Supervised Clustering

- Domain knowledge
  - Partial label information is given
  - Apply some constraints (must-links and cannot-links)

- Approaches
  - Search-based Semi-Supervised Clustering
    - Alter the clustering algorithm using the constraints
  - Similarity-based Semi-Supervised Clustering
    - Alter the similarity measure based on the constraints
  - Combination of both

What is SSL ?



labeled data
{ (x_i,y_i) }$_i$ i=1,…,N
unlabeled data
{ x_j }$_j$ j=N+1,…,N+M
Supervised Learning
estimate f such that f(x) = y
other view: describe Pr(y|x)
use both
Unsupervised Learning
characterize Pr(x)
e.g. by clustering or PCA
Semi-Supervised Learning
same goal as supervised learning: Pr(y|x)
additional data { x_j }$_j$ j=N+1,…,N+M

What is SSL ?

Goal:

Using both labeled and unlabeled data to build better classifiers (than using labeled data alone).

Notation:

- input x, label y
- classifier $f : \mathcal{X} \mapsto \mathcal{Y}$
- labeled data $(X_l, Y_l) = \{(x_1, y_1), \ldots, (x_l, y_l)\}$
- unlabeled data $X_u = \{x_{l+1}, \ldots, x_n\}$
- usually $n \gg l$

Semi-supervised
self and co-training

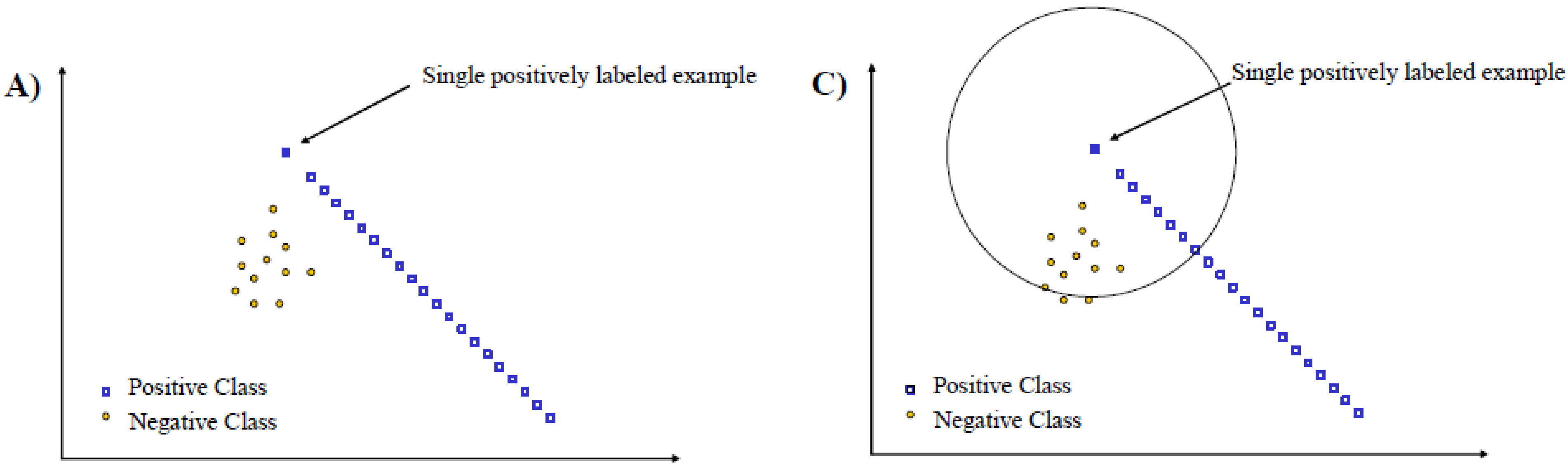
Self-training

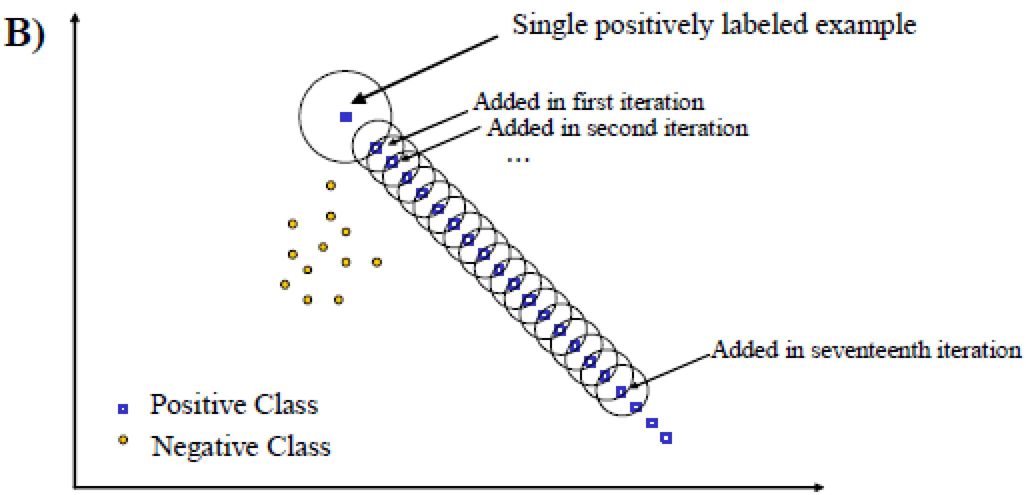
Algorithm: Self-training

1. Pick your favorite classification method. Train a classifier f from (X_l, Y_l).
2. Use f to classify all unlabeled items $x \in X_u$.
3. Pick x^* with the highest confidence, add $(x^*, f(x^*))$ to labeled data.
4. Repeat.

The training needs a point to stop its processing, called “STOPPING CRITERIA”

Self-training: (2-class Toy example)

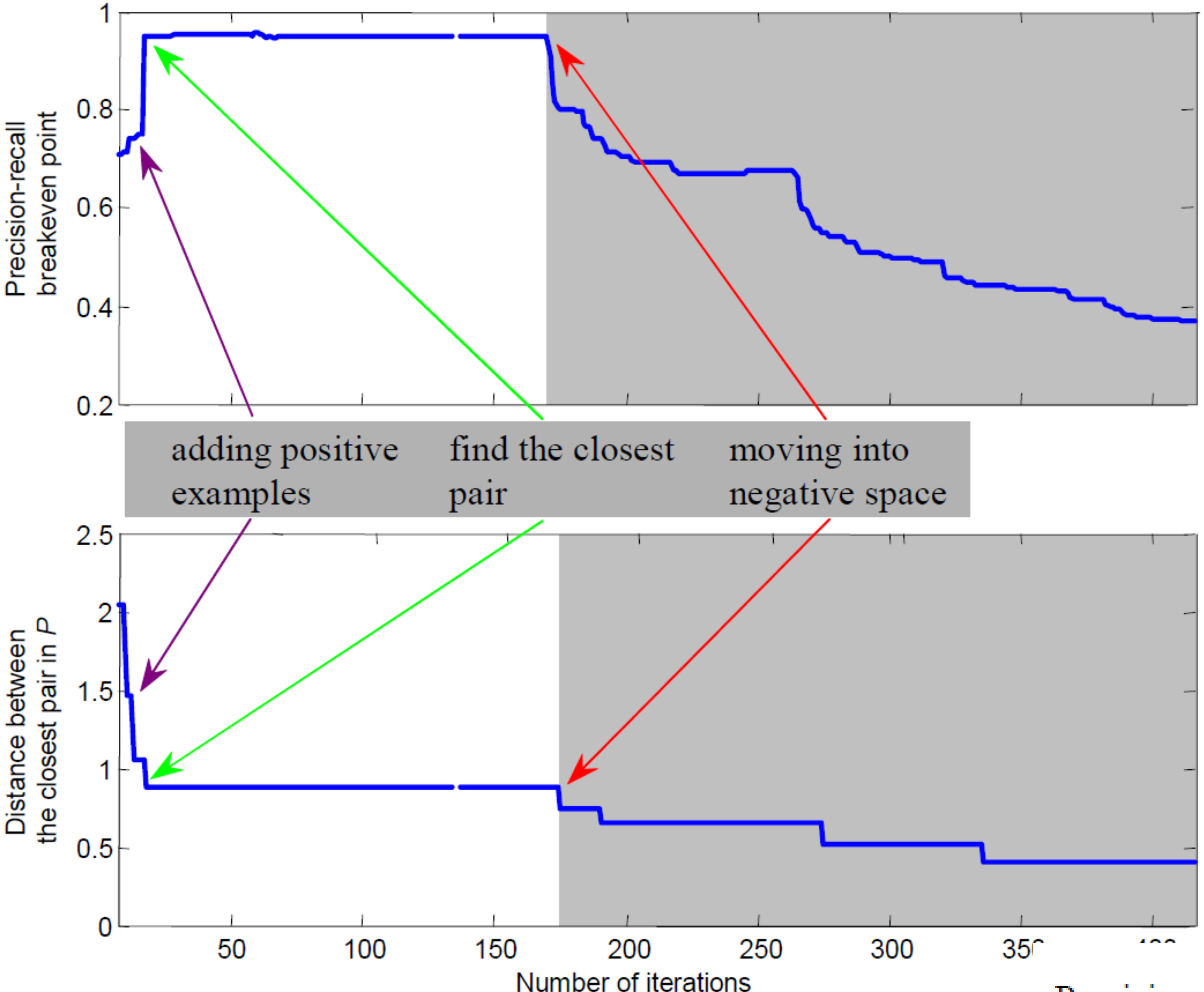




- **Ref: Li Wei and Eamonn Keogh (2006) Semi-Supervised Time Series Classification. SIGKDD 2006.**

A) A simple two-class dataset. B) The chaining effect of semi-supervised learning: a positive example is labeled which helps labeling other positive examples and so on. Eventually all positive examples are correctly classified. **C) If we simply put the** seventeen nearest neighbors of the single labeled example to the positive class, we would wrongly include many negative examples into the positive class

Self-training: (stopping criteria)

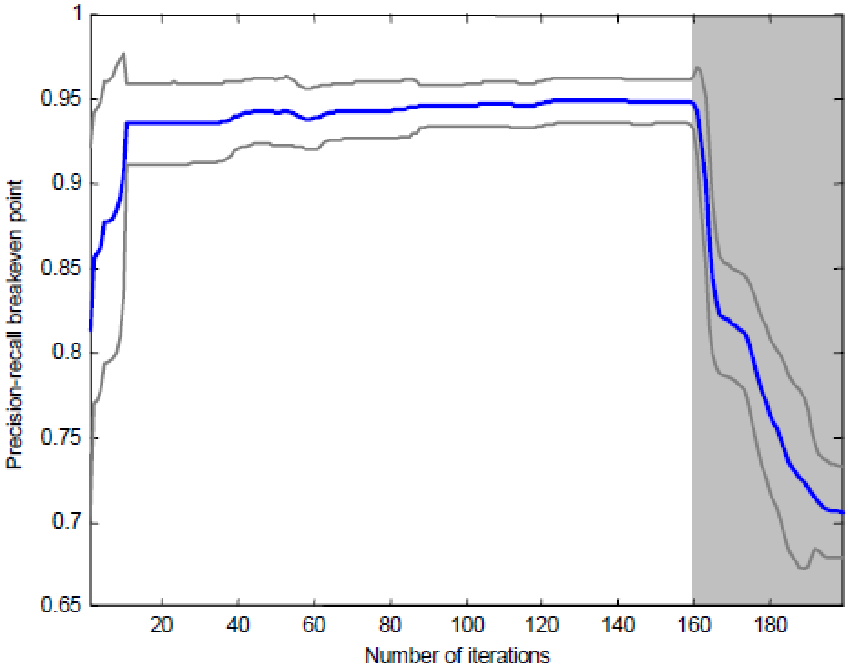


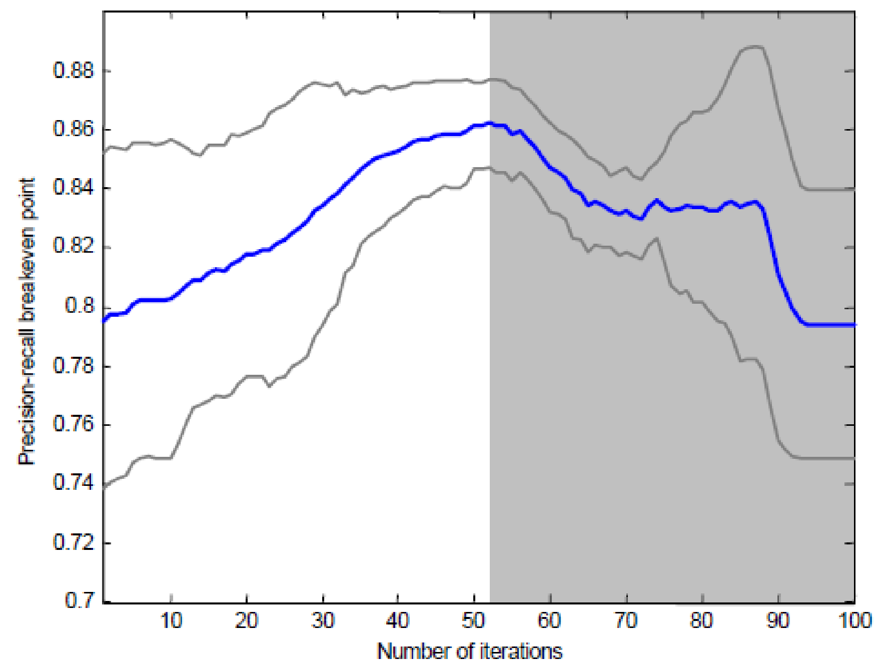
$$\text{Precision} = \frac{\#\text{ of correct positive predictions}}{\#\text{ of positive predictions}}$$

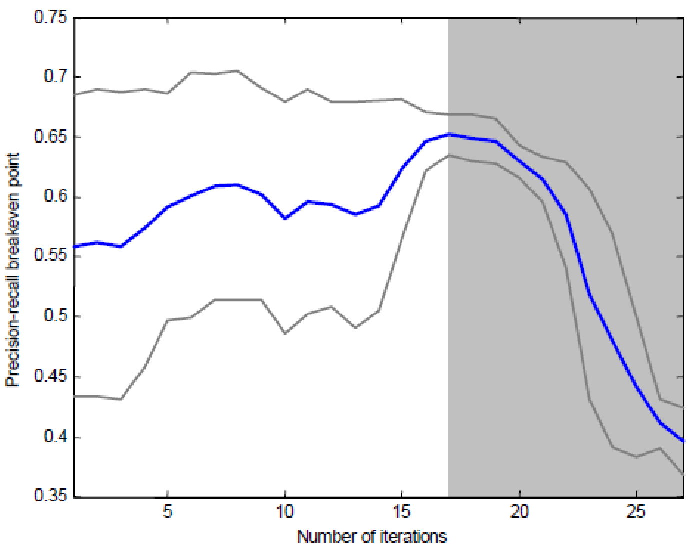
$$\text{Recall} = \frac{\#\text{ of correct positive predictions}}{\#\text{ of positive examples}}$$

- **Ref: Li Wei and Eamonn Keogh (2006) Semi-Supervised Time Series Classification. SIGKDD 2006.**

Self-training: (stopping criteria, some more results)







- **Ref: Li Wei and Eamonn Keogh (2006) Semi-Supervised Time Series Classification. SIGKDD 2006.**

Pros and cons of self-training

Pros

- Simple
- Applies to almost all existing classifiers

Cons

- Mistakes reinforce/strengthen themselves. Heuristics against pitfalls
  - ‘Un-label’ a training point if its classification confidence drops below a threshold
  - Randomly perturb/disturb learning parameters

- Can’t say too much

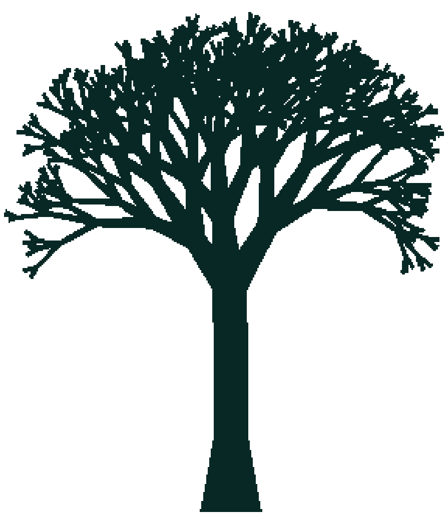
CO-TRAINING

Two, out of different views of an item: image and HTML text

Answer:

Saturn is the second largest planet in our solar system and the sixth in distance from the Sun. It is mostly gaseous with large "rings" of ice and rocks. It has more than 60 moons, the inner ones small and within the ring system. The outer hydrogen atmosphere is very cold, as low as -200 °C. But the small rocky core is heated by presssure to over 11700°C.

Saturn cannot support life because the dense liquid hydrogen has incredibly high pressure at greater depths and may form a mantle of solid hydrogen



Trees are an important part of our daily lives. They also absorb carbon dioxide (a greenhouse gas) and give us oxygen to breathe. Trees make our environment beautiful with their different colours, flowers and shapes and they provide us with shade and relief from the sun's heat and harmful rays. Trees help absorb the rain and help stabilize the weather. Trees are very important to us!

CO-TRAINING: Feature split

Each item is represented by two kinds of features

$x = [x^{(1)}; x^{(2)}]$

- $x^{(1)}$ = image features
- $x^{(2)}$ = web page text
- This is a natural feature split (or multiple views)

Co-training idea:

- Train an image classifier and a text classifier
- The two classifiers teach each other

- **Ref: Semi-supervised classification. Xiaojin Zhu. Univ. Wisconsin-Madison**

CO-TRAINING: Algorithm

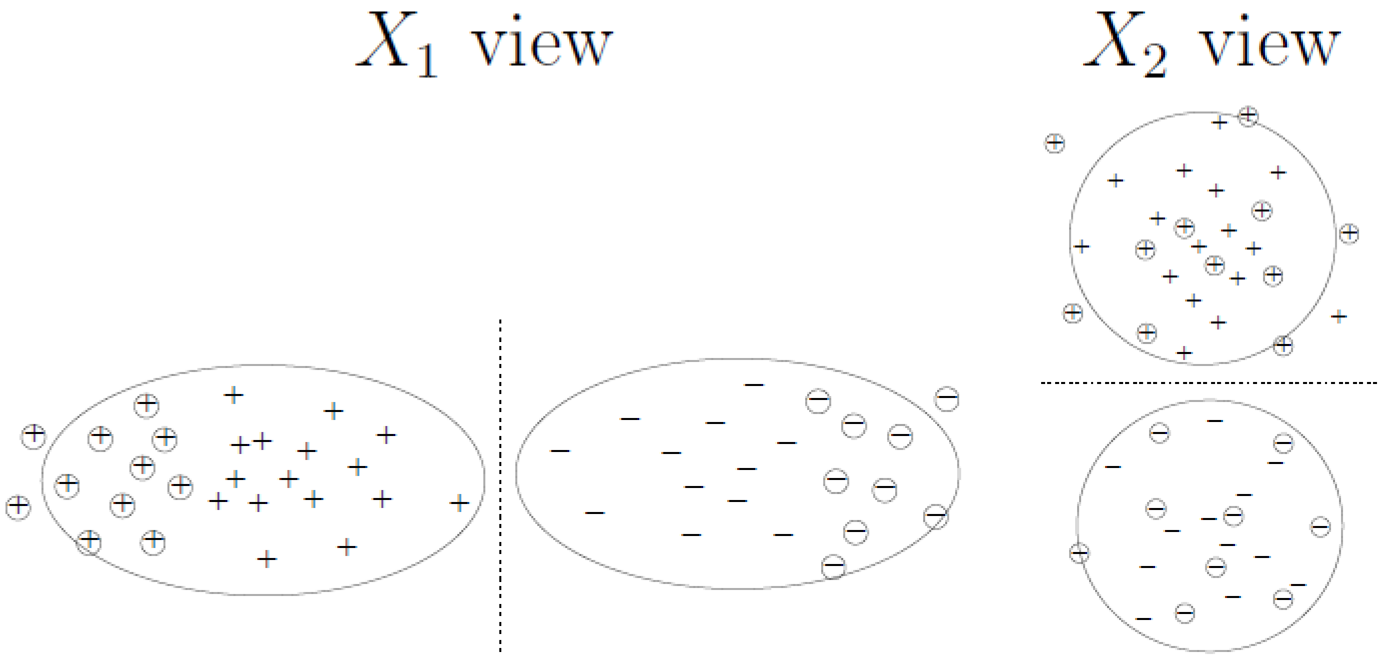
Algorithm: Co-training

1. Train two classifiers: $f^{(1)}$ from $(X_l^{(1)}, Y_l)$, $f^{(2)}$ from $(X_l^{(2)}, Y_l)$.
2. Classify X_u with $f^{(1)}$ and $f^{(2)}$ separately.
3. Add $f^{(1)}$'s k-most-confident $(x, f^{(1)}(x))$ to $f^{(2)}$'s labeled data.
4. Add $f^{(2)}$'s k-most-confident $(x, f^{(2)}(x))$ to $f^{(1)}$'s labeled data.
5. Repeat.

CO-TRAINING: Assumptions

Co-training assumes that

- feature split $x = [x^{(1)}; x^{(2)}]$ exists
- $x^{(1)}$ or $x^{(2)}$ alone is sufficient to train a good classifier
- $x^{(1)}$ and $x^{(2)}$ are conditionally independent given the class



Pros and cons of co-training

Pros

- Simple. Applies to almost all existing classifiers
- Less sensitive to mistakes

Cons

- Feature split may not exist
- Models using BOTH features should do better

Variants of co-training

Co-EM: add all, not just top *k*

- Each classifier probabilistically label X_u
- Add (x, y) with weight P(y|x)

Single-view: fake feature split

- create random, artificial feature split
- apply co-training

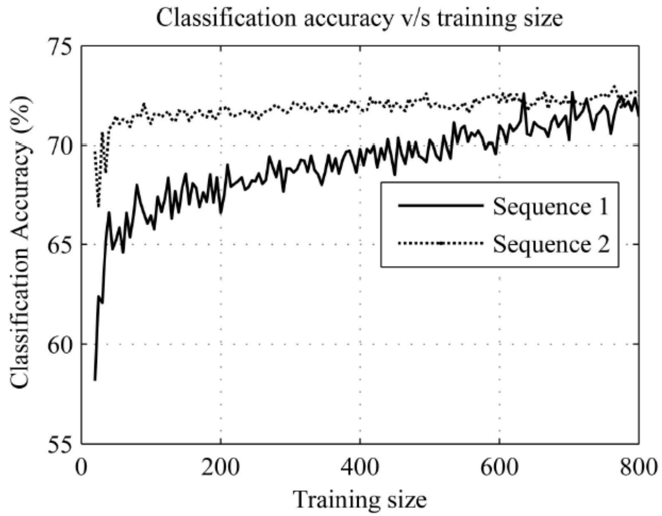
Single-view: agreement among multiple classifiers

- train multiple classifiers of different types
- classify unlabeled data with all classifiers
- add majority vote label

Results and Discussion

Ref: Ajay J. Joshi and Nikolaos P. Papanikolopoulos, “Learning to Detect Moving Shadows in Dynamic Environments”, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 30, NO. 11, NOVEMBER 2008

Results: co-training



Row	Feature Split	Rounds of Co-training			
	$[Set1]$ $[Set2]$	20	50	100	200
1	$[f_1]$ $[f_2\ f_3\ f_4]$	61.96	60.56	59.91	60.20
2	$[f_2]$ $[f_1\ f_3\ f_4]$	74.27	73.48	73.28	72.89
3	$[f_3]$ $[f_1\ f_2\ f_4]$	57.00	57.89	59.21	60.41
4	$[f_4]$ $[f_1\ f_2\ f_3]$	**75.16**	**75.22**	**74.62**	**74.72**
5	$[f_1\ f_2]$ $[f_3\ f_4]$	**74.95**	**75.50**	**74.95**	**75.30**
6	$[f_1\ f_3]$ $[f_2\ f_4]$	61.72	60.95	61.22	61.63
7	$[f_1\ f_4]$ $[f_2\ f_3]$	73.59	73.09	73.01	73.24
8	No Co-training	66.50			

(a)

Row	Feature Split	Rounds of Co-training			
	$[Set1]$ $[Set2]$	20	50	100	200
1	$[f_1]$ $[f_2\ f_3\ f_4]$	70.15	75.34	69.88	69.81
2	$[f_2]$ $[f_1\ f_3\ f_4]$	74.47	74.64	74.54	74.43
3	$[f_3]$ $[f_1\ f_2\ f_4]$	71.65	71.94	71.47	70.58
4	$[f_4]$ $[f_1\ f_2\ f_3]$	**76.63**	**76.82**	**76.55**	**76.26**
5	$[f_1\ f_2]$ $[f_3\ f_4]$	**76.06**	**75.86**	**75.52**	**75.19**
6	$[f_1\ f_3]$ $[f_2\ f_4]$	71.83	71.41	71.19	70.92
7	$[f_1\ f_4]$ $[f_2\ f_3]$	75.65	75.62	75.38	75.22
8	No Co-training	72.34			

(b)

Performance with two data sets

Conclusions

- Labeled data are expensive and limited that motivates to use unlabeled data which can significantly help to improve classification accuracy along with the labeled data
- Combining generative probabilistic models leads to natural use of unlabeled data
  - Unlabeled data don’t always lead to performance gain
  - Depend on whether the generative model is correct or not
- Co-training assumes that there are two redundant and conditionally independent feature sets
  - In practice there is often no natural split of features, and hence random splits can help as well