LAMDA
Learning And Mining from DatA
http://lamda.nju.edu.cn

南京大學

When Semi-Supervised Learning Meets Ensemble Learning

Zhi-Hua Zhou

http://cs.nju.edu.cn/zhouzh/
Email: zhouzh@nju.edu.cn

LAMDA Group
National Key Laboratory for Novel Software Technology,
Nanjing University, China

The presentation involves some joint work with :

Ming Li

Wei Wang

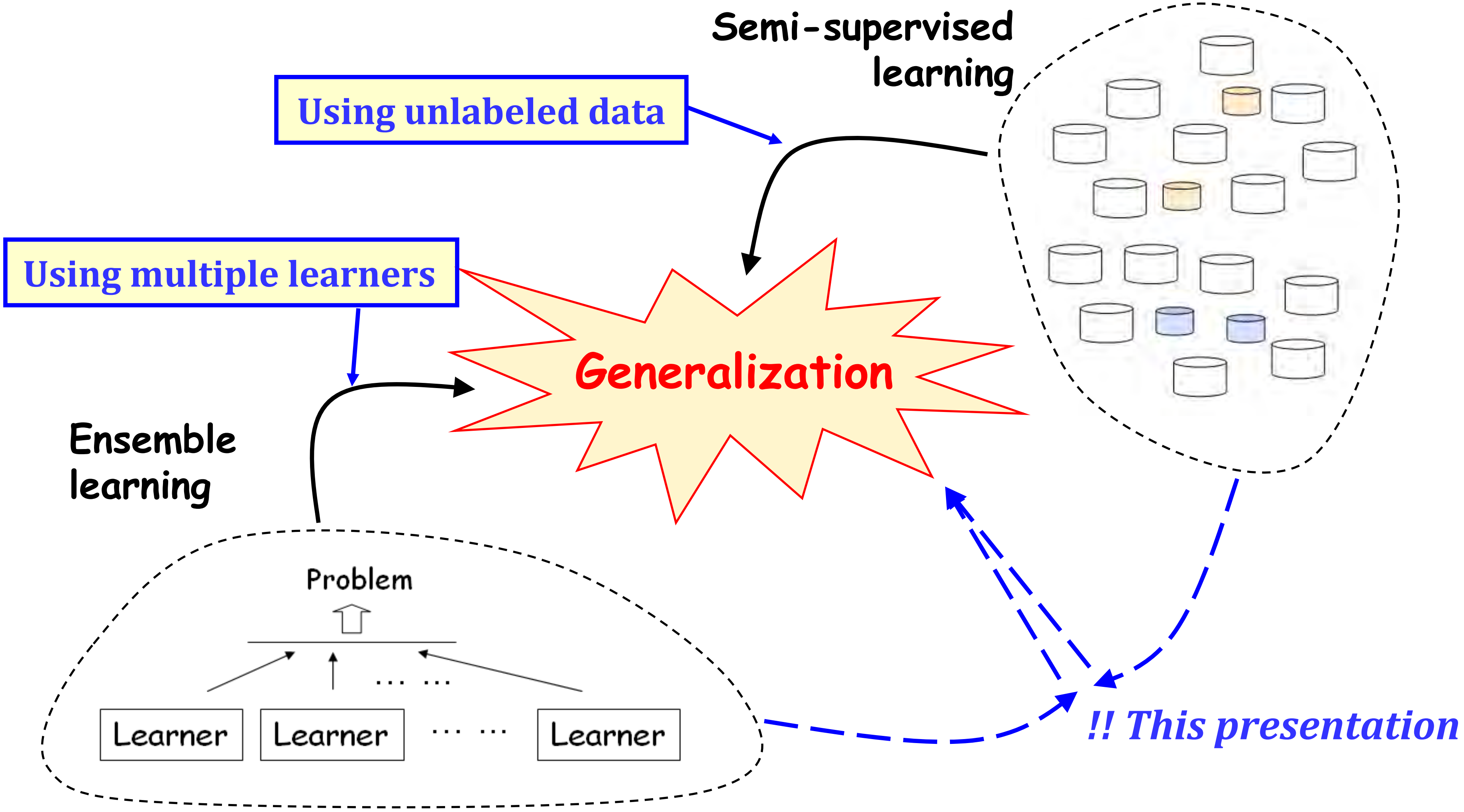
Qiang Yang

Min-Ling Zhang

De-Chuan Zhan

… …

One Goal, Two Paradigms



Semi-supervised learning

Using unlabeled data

Using multiple learners

Generalization

Ensemble learning

!! This presentation

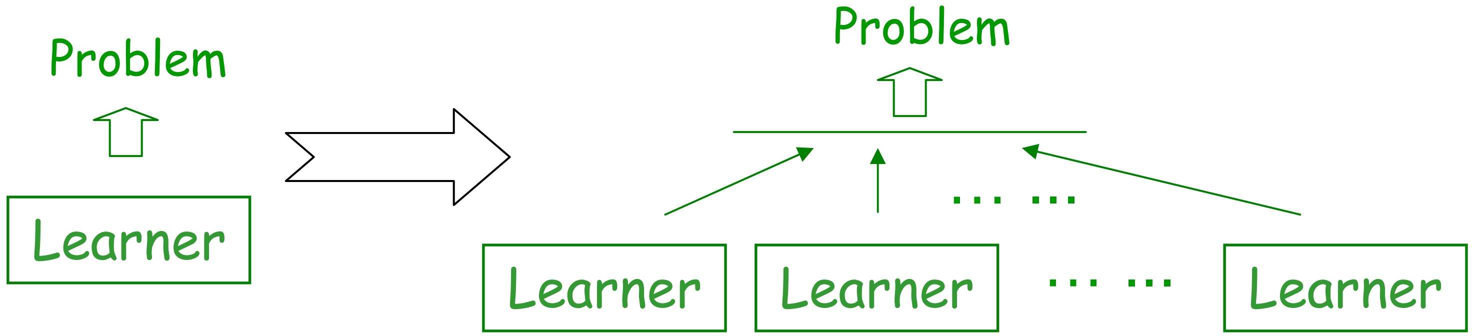
Outline

- Ensemble Learning
- Semi-Supervised Learning
- Classifier Combination vs. Unlabeled Data

What’s ensemble learning?

Ensemble learning is a machine learning paradigm where multiple (homogenous/heterogeneous) individual learners are trained for the same problem

e.g. neural network ensemble, decision tree ensemble, etc.



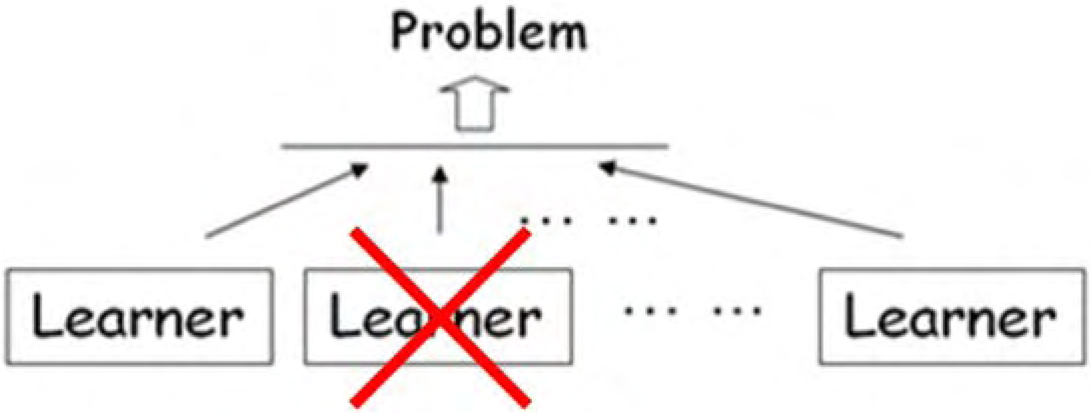
Many ensemble methods

- Parallel methods
  - Bagging [L. Breiman, MLJ96]
  - Random Subspace [T. K . Ho, TPAMI98]
  - Random Forests [L. Breiman, MLJ01]
  - … …
- Sequential methods
  - AdaBoost [Y. Freund & R. Schapire, JCSS97]
  - Arc-x4 [L. Breiman, AnnStat98]
  - LPBoost [A. Demiriz et al., MLJ06]
  - … …

Selective ensemble

Many Could be Better Than All:

When a number of base learners are available, …, ensembling **many** of the base learners may be better than ensembling **all** of them [Z.-H. Zhou et al., IJCAI'01 & AIJ02]



Theoretical foundations

Abundant studies on theoretical properties of ensemble methods

Appeared/ing in many leading statistical journals, e.g. *Annals of Statistics*

√ Agreement: Different ensemble methods may have different foundations

Many mysteries

Diversity among the base learners is (possibly) the key of ensembles

$$E = \overline{E} - \overline{A}$$ [A. Krogh & J. Vedelsby, NIPS'94]

The more accurate and the more diverse, the better

but, what is "diversity"? [L.I. Kuncheva & C.J. Whitaker, MLJ03]

Many mysteries (con't)

Even for some theory-intrigued methods, … still mysteries

E.g., Why AdaBoost does not overfit?

- **Margin !** [R.E. Schapire et al., AnnStat98]
- **No!** [L. Breiman, NCJ99]

  (contrary evidence: minimal margin)
- **Wait …** [L. Reyzin & R.E. Schapire, ICML'06 best paper]

  (minimal Margin ?? Margin distribution)
- **One more support** [L. Wang et al., COLT'08]

For the whole story see:

Z.-H. Zhou & Y. Yu, **AdaBoost**. In: X. Wu and V. Kumar eds. The Top Ten Algorithms in Data Mining, Boca Raton, FL: Chapman & Hall, 2009

Great success of ensemble methods

□ **KDDCup'05**: **all awards** (“Precision Award”, “Performance Award”, “Creativity Award”) **for** “*An* ***ensemble*** *search based method …* ”

□ **KDDCup'06**: **1st place of Task1 for** “*Modifying* ***Boosted Trees*** *to …* ”; **1st place of Task2 & 2nd place of Task1 for** “***Voting*** *… by means of a* ***Classifier Committee***”

□ **KDD Time-series Classification Challenge 2007**: **1st place for** “*…* ***Decision Forests*** *and …*”

Great success of ensemble methods (con't)

- □ **KDDCup'08**: **1st place of Challenge1 for a method using Bagging;** **1st place of Challenge2 for "…** *Using an* ***Ensemble Method*** **"**

- □ **KDDCup'09**: **1st place of Fast Track for "*Ensemble* … "; 2nd place of Fast Track for "…** ***bagging*** **…** ***boosting tree models*** **…", 1st place of Slow Track for "*Boosting*** *with classification trees and shrinkage*"**; 2nd place of Slow Track for "*Stochastic Gradient Boosting*"**

- □ … …

Great success of ensemble methods (con't)

- □ **Netflix Prize**:
  - ✓ **2007 Progress Prize Winner**: ***Ensemble***
  - ✓ **2008 Progress Prize Winner**: ***Ensemble***
  - ✓ **2009 $1 Million Grand Prize Winner**: ***Ensemble !!***
- □ "Top 10 Data Mining Algorithms" (ICDM'06): AdaBoost
- □ **Application to almost all areas**
- □ … …

Recently, very few papers in

top machine learning conferences

Why?

Easier tasks finished

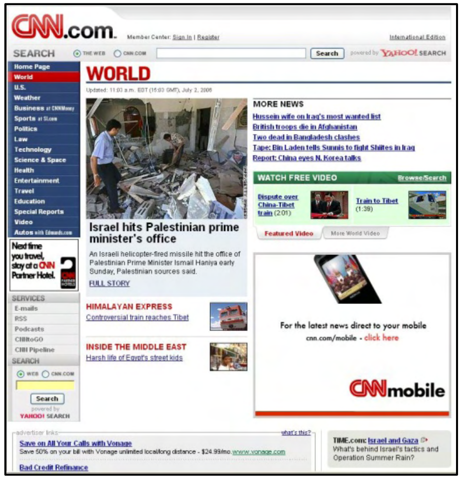
New challenges needed

Outline

- Ensemble Learning
- **Semi-Supervised Learning**
- Classifier Combination vs. Unlabeled Data

Labeled vs. Unlabeled

In many practical applications, **unlabeled** training examples are readily available but labeled ones are fairly expensive to obtain because labeling the unlabeled examples requires human effort



CNN.com Member Center: Sign In | Register International Edition

SEARCH ◉ THE WEB ○ CNN.COM Search powered by YAHOO! SEARCH

Home Page
World
U.S.
Weather
Business at CNNMoney
Sports at SI.com
Politics
Law
Technology
Science & Space
Health
Entertainment
Travel
Education
Special Reports
Video
Autos with Edmunds.com

Next time you travel, stay at a CNN Partner Hotel.

SERVICES
E-mails
RSS
Podcasts
CNNtoGO
CNN Pipeline

SEARCH
◉ WEB ○ CNN.COM
Search
powered by YAHOO! SEARCH

WORLD

Updated: 11:03 a.m. EDT (15:03 GMT), July 2, 2006

Israel hits Palestinian prime minister's office

An Israeli helicopter-fired missile hit the office of Palestinian Prime Minister Ismail Haniya early Sunday, Palestinian sources said.

FULL STORY

HIMALAYAN EXPRESS
Controversial train reaches Tibet

INSIDE THE MIDDLE EAST
Harsh life of Egypt's street kids

MORE NEWS
Hussein wife on Iraq's most wanted list
British troops die in Afghanistan
Two dead in Bangladesh clashes
Tape: Bin Laden tells Sunnis to fight Shiites in Iraq
Report: China eyes N. Korea talks

WATCH FREE VIDEO Browse/Search
Dispute over China-Tibet train (2:01)
Train to Tibet (1:39)
Featured Video | More World Video

For the latest news direct to your mobile
cnn.com/mobile - click here
CNNmobile

advertiser links what's this?
Save on All Your Calls with Vonage
Save 50% on your bill with Vonage unlimited local/long distance - $24.99/mo. www.vonage.com
Bad Credit Refinance

TIME.com: Israel and Gaza
What's behind Israel's tactics and Operation Summer Rain?

class = "*war*"

?

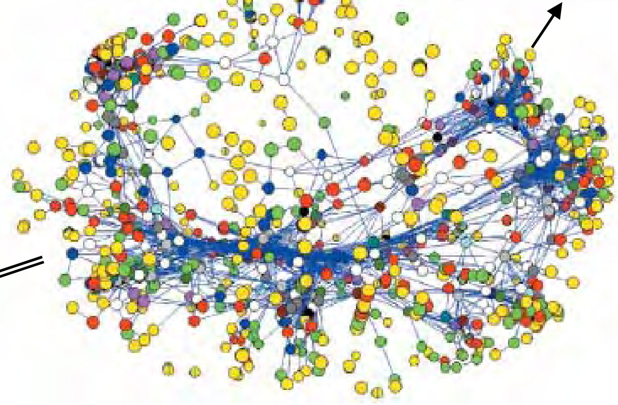
(almost) infinite number of web pages on the Internet

YAHOO! SHOPPING AUCTIONS

Books & Comics

Auctions > Books & Comics

Search Books & Comics Advanced Search Sell Stuff



SSL: Why unlabeled data can be helpful?

Suppose the data is well-modeled by a mixture density:

$$f\left(x|\theta\right)=\sum_{l=1}^{L}\alpha_l f\left(x|\theta_l\right) \quad \text{where} \sum_{l=1}^{L}\alpha_l=1 \text{ and } \theta=\{\theta_l\}$$

The class labels are viewed as random quantities and are assumed chosen conditioned on the selected mixture component $m_i \in \{1,2,\ldots,L\}$ and possibly on the feature value, i.e. according to the probabilities $\mathrm{P}[c_i | x_i, m_i]$

Thus, the optimal classification rule for this model is the MAP rule:

$$S\left(x\right)=\arg\max_k \sum_j \mathrm{P}\left[c_i=k|m_i=j,x_i\right]\underline{\mathrm{P}\left[m_i=j|x_i\right]}$$

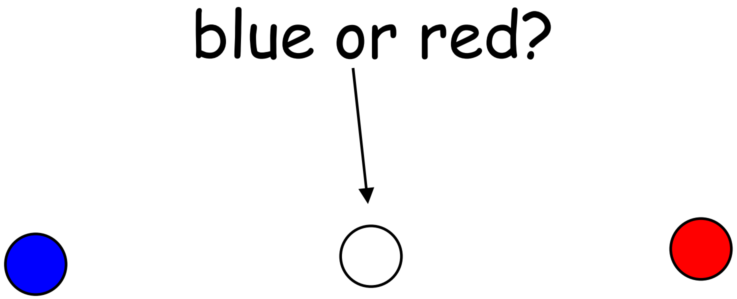
$$\text{where } \mathrm{P}\left[m_i=j|x_i\right]=\frac{\alpha_j f\left(x_i|\theta_j\right)}{\sum_{l=1}^{L}\alpha_l f\left(x_i|\theta_l\right)}$$

unlabeled examples can be used to help estimate this term

[D.J. Miller & H.S. Uyar, NIPS'96]

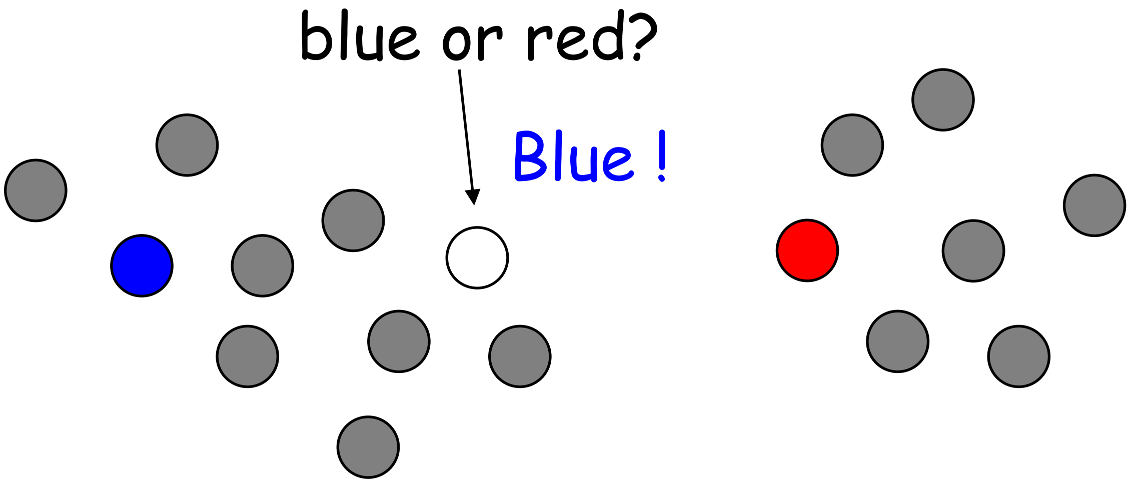
SSL: Why unlabeled data can be helpful? (con't)

Intuitively,



SSL: Why unlabeled data can be helpful? (con't)

Intuitively,



SSL: Representative approaches

- ✓ **Generative methods**

  Using a generative model for the classifier and employing EM to model the label estimation or parameter estimation process [Miller & Uyar, NIPS'96; Nigam et al., MLJ00; Fujino et al., AAAI'05; etc.]

- ✓ **S3VMs (Semi-Supervised SVMs)**

  Using unlabeled data to adjust the decision boundary such that it goes through the less dense region [Joachims, ICML'99; Chapelle & Zien, AISTATS'05; Collobert et al., ICML'06; etc.]

- ✓ **Graph-based methods**

- ✓ **Disagreement-based methods**

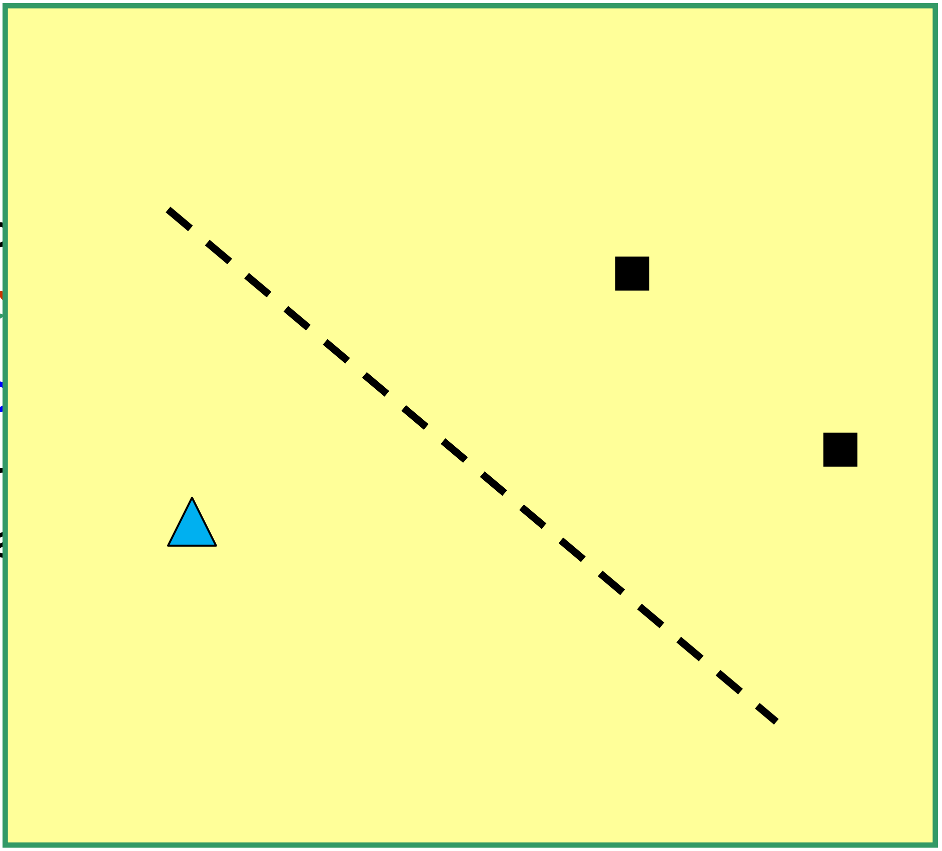
SSL: Representative approaches

- ✓ Generative methods

  Using a generative model for t
  model the label estimation or p
  & Uyar, NIPS'96; Nigam et al., ML

- ✓ S3VMs (Semi-Supervised

  Using unlabeled data to adjust
  goes through the less dense re
  Zien, AISTATS'05; Collobert et a



- ✓ Graph-based methods

- ✓ Disagreement-based methods

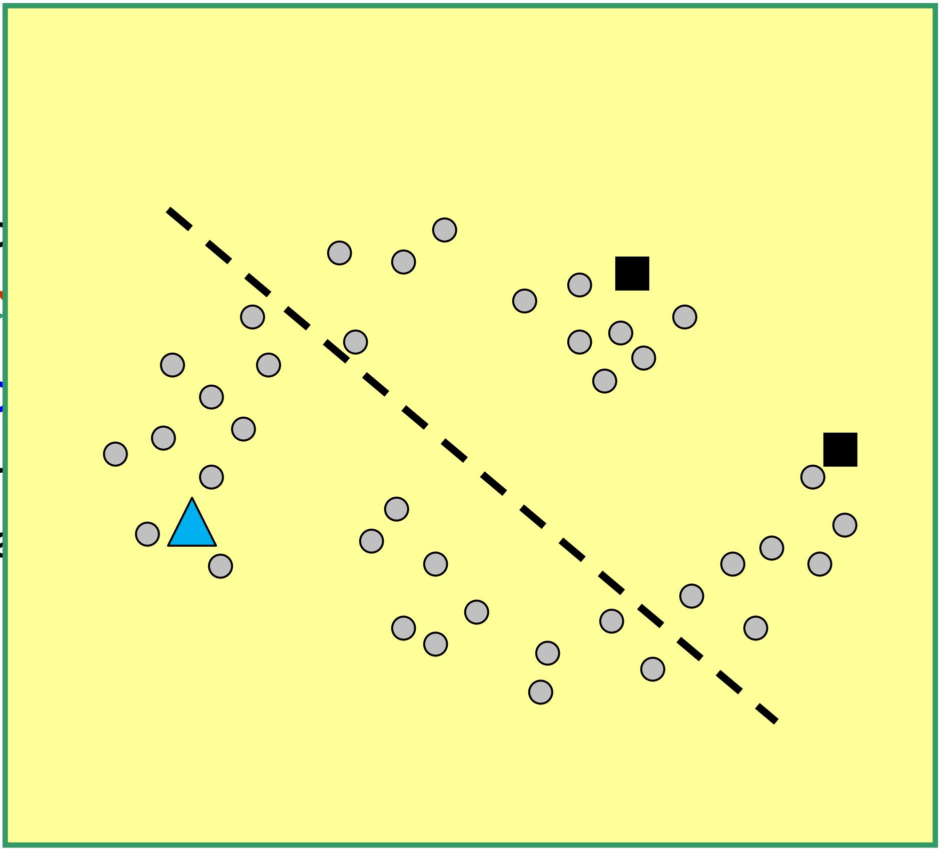
SSL: Representative approaches

- ✓ Generative methods

  Using a generative model for t
  model the label estimation or p
  & Uyar, NIPS'96; Nigam et al., ML

- ✓ S3VMs (Semi-Supervised

  Using unlabeled data to adjust
  goes through the less dense re
  Zien, AISTATS'05; Collobert et a



- ✓ Graph-based methods

- ✓ Disagreement-based methods

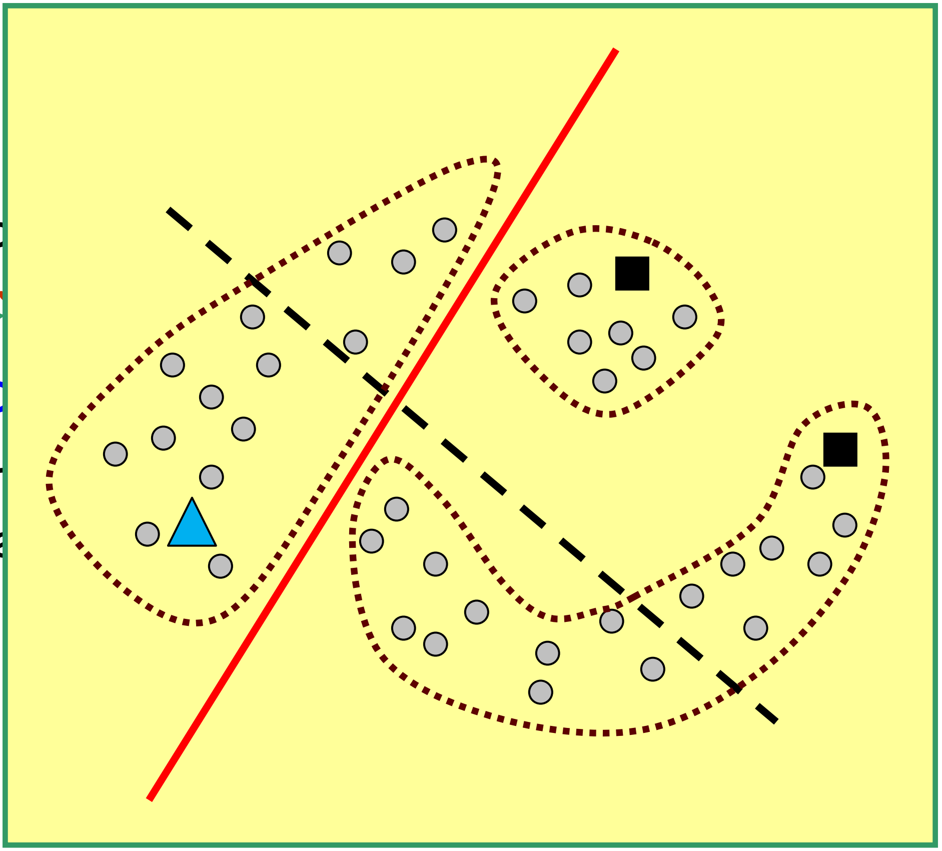
SSL: Representative approaches

- ✓ Generative methods

  Using a generative model for t
  model the label estimation or p
  & Uyar, NIPS'96; Nigam et al., ML

- ✓ S3VMs (Semi-Supervised

  Using unlabeled data to adjust
  goes through the less dense re
  Zien, AISTATS'05; Collobert et a



- ✓ Graph-based methods

- ✓ Disagreement-based methods

SSL: Representative approaches

- ✓ **Generative methods**

  Using a generative model for the classifier and employing EM to model the label estimation or parameter estimation process [Miller & Uyar, NIPS'96; Nigam et al., MLJ00; Fujino et al., AAAI'05; etc.]

- ✓ **S3VMs (Semi-Supervised SVMs)**

  Using unlabeled data to adjust the decision boundary such that it goes through the less dense region [Joachims, ICML'99; Chapelle & Zien, AISTATS'05; Collobert et al., ICML'06; etc.]

- ✓ **Graph-based methods**

  Using unlabeled data to regularize the learning process via graph regularization [Blum & Chawla, ICML'01; Belkin & Niyogi, MLJ04; Zhou et al., NIPS'04; etc.]

- ✓ **Disagreement-based methods**

SSL: Representative approaches

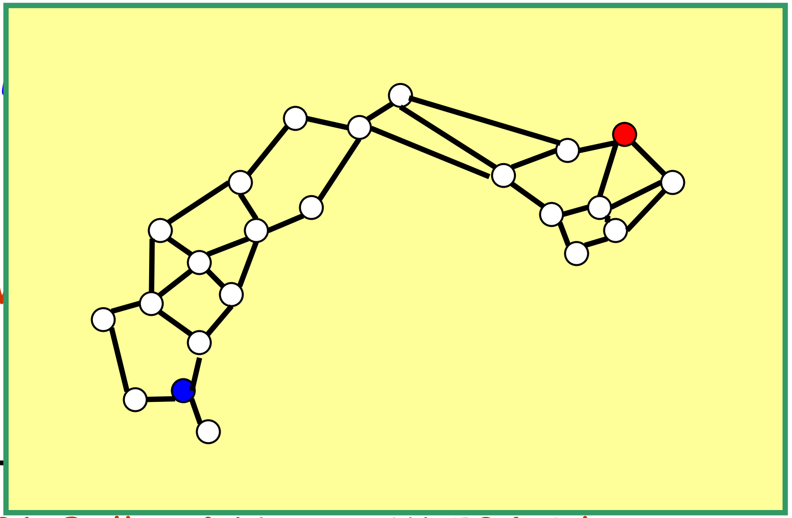
- ✓ Generative methods

  Using a generative model for the classifier and employing EM to model the label estimation or parameter estimation process [Miller & Uyar, NIPS'96; Nigam et al., MLJ00; Fujino et al., AAAI'05; etc.]

- ✓ S3VMs (Semi-Supervised SV

  Using unlabeled data to adjust the
  goes through the less dense region
  Zien, AISTATS'05; Collobert et al., ICM

- ✓ Graph-based methods



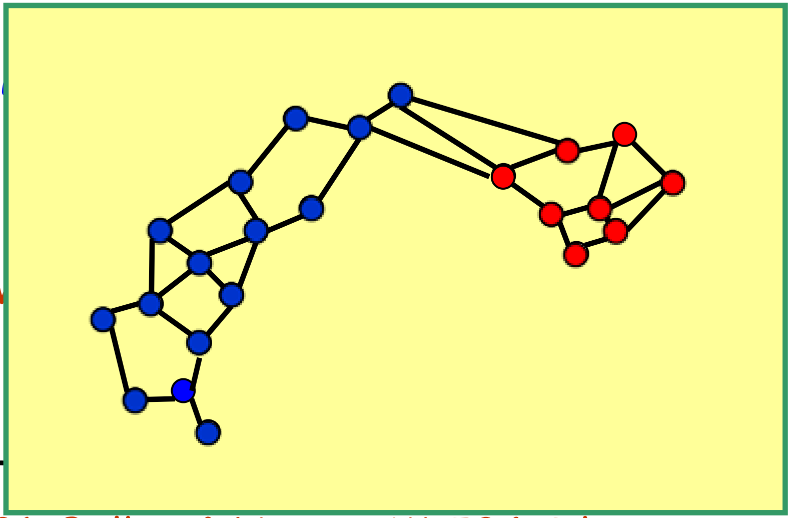
  Using unlabeled data to regularize t
  regularization [Blum & Chawla, ICML'01; Belkin & Niyogi, MLJ04; Zhou et al., NIPS'04; etc.]

- ✓ Disagreement-based methods

SSL: Representative approaches

- ✓ **Generative methods**

  Using a generative model for the classifier and employing EM to model the label estimation or parameter estimation process [Miller & Uyar, NIPS'96; Nigam et al., MLJ00; Fujino et al., AAAI'05; etc.]



- ✓ **S3VMs (Semi-Supervised SV**

  Using unlabeled data to adjust the goes through the less dense region Zien, AISTATS'05; Collobert et al., ICM

- ✓ **Graph-based methods**

  Using unlabeled data to regularize t regularization [Blum & Chawla, ICML'01; Belkin & Niyogi, MLJ04; Zhou et al., NIPS'04; etc.]

- ✓ **Disagreement-based methods**

SSL: Representative approaches

✓ Generative methods

✓ S3VMs (Semi-Supervised SVMs)

✓ Graph-based methods

✓ Disagreement-based methods

multiple learners are trained for the task and the disagreements among the learners are exploited during the SSL process [Blum & Mitchell, COLT'98; Goldman & Zhou, ICML'00; Zhou & Li, TKDE05; etc.]

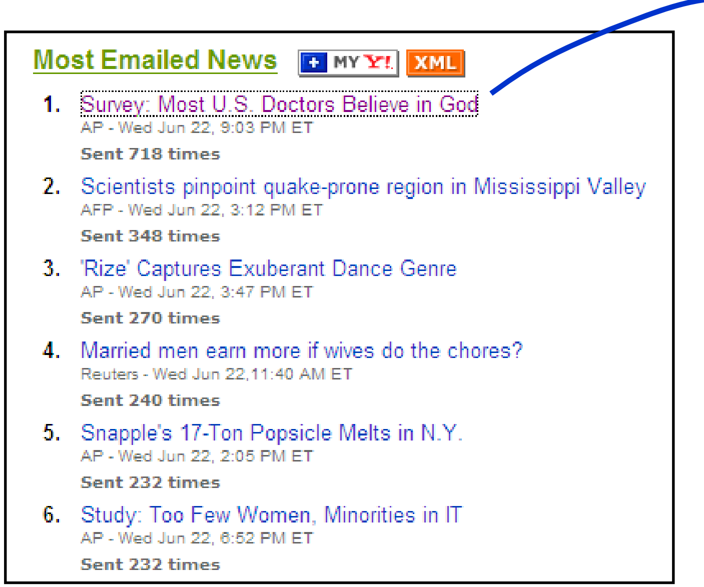
SSL reviews:

- Chapelle et al., eds. Semi-Supervised Learning, MIT Press, 2006
- Zhu, Semi-Supervise Learning Literature Survey, 2006
- Zhou & Li, Semi-supervised learning by disagreement, KAIS, 2009

Co-training

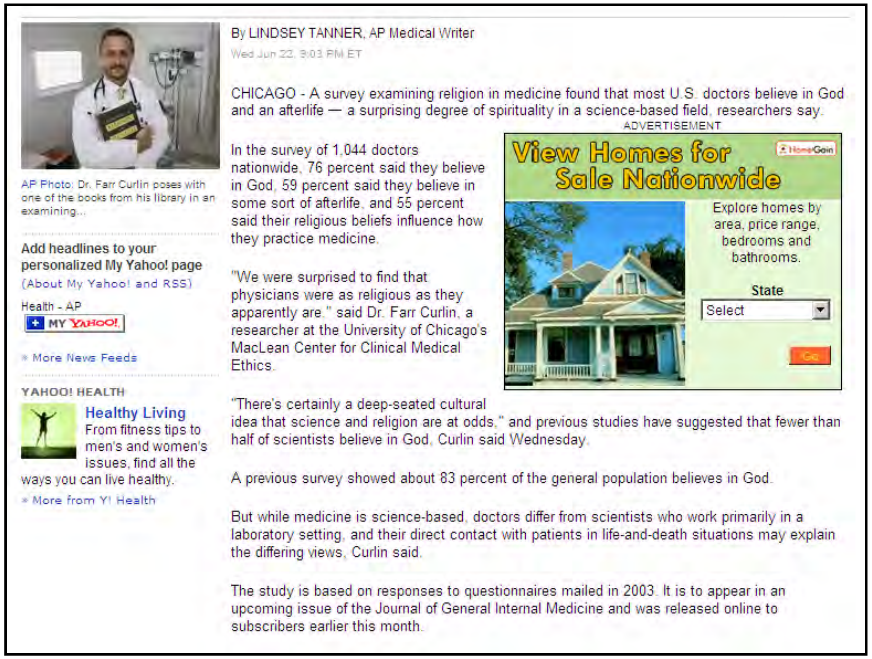
In some applications, there are two sufficient and redundant views, i.e. two attribute sets each of which is sufficient for learning and conditionally independent to the other given the class label

e.g. two views for web page classification: 1) the text appearing on the page itself, and 2) the anchor text attached to hyperlinks pointing to this page, from other pages



Most Emailed News + MY Y! XML

1. Survey: Most U.S. Doctors Believe in God
   AP - Wed Jun 22, 9:03 PM ET
   Sent 718 times
2. Scientists pinpoint quake-prone region in Mississippi Valley
   AFP - Wed Jun 22, 3:12 PM ET
   Sent 348 times
3. 'Rize' Captures Exuberant Dance Genre
   AP - Wed Jun 22, 3:47 PM ET
   Sent 270 times
4. Married men earn more if wives do the chores?
   Reuters - Wed Jun 22, 11:40 AM ET
   Sent 240 times
5. Snapple's 17-Ton Popsicle Melts in N.Y.
   AP - Wed Jun 22, 2:05 PM ET
   Sent 232 times
6. Study: Too Few Women, Minorities in IT
   AP - Wed Jun 22, 6:52 PM ET
   Sent 232 times



AP Photo: Dr. Farr Curlin poses with one of the books from his library in an examining...

Add headlines to your personalized My Yahoo! page
(About My Yahoo! and RSS)

Health - AP
+ MY YAHOO!

» More News Feeds

YAHOO! HEALTH

Healthy Living
From fitness tips to men's and women's issues, find all the ways you can live healthy.

» More from Y! Health

By LINDSEY TANNER, AP Medical Writer
Wed Jun 22, 9:03 PM ET

CHICAGO - A survey examining religion in medicine found that most U.S. doctors believe in God and an afterlife — a surprising degree of spirituality in a science-based field, researchers say.

In the survey of 1,044 doctors nationwide, 76 percent said they believe in God, 59 percent said they believe in some sort of afterlife, and 55 percent said their religious beliefs influence how they practice medicine.

"We were surprised to find that physicians were as religious as they apparently are," said Dr. Farr Curlin, a researcher at the University of Chicago's MacLean Center for Clinical Medical Ethics.

"There's certainly a deep-seated cultural idea that science and religion are at odds," and previous studies have suggested that fewer than half of scientists believe in God, Curlin said Wednesday.

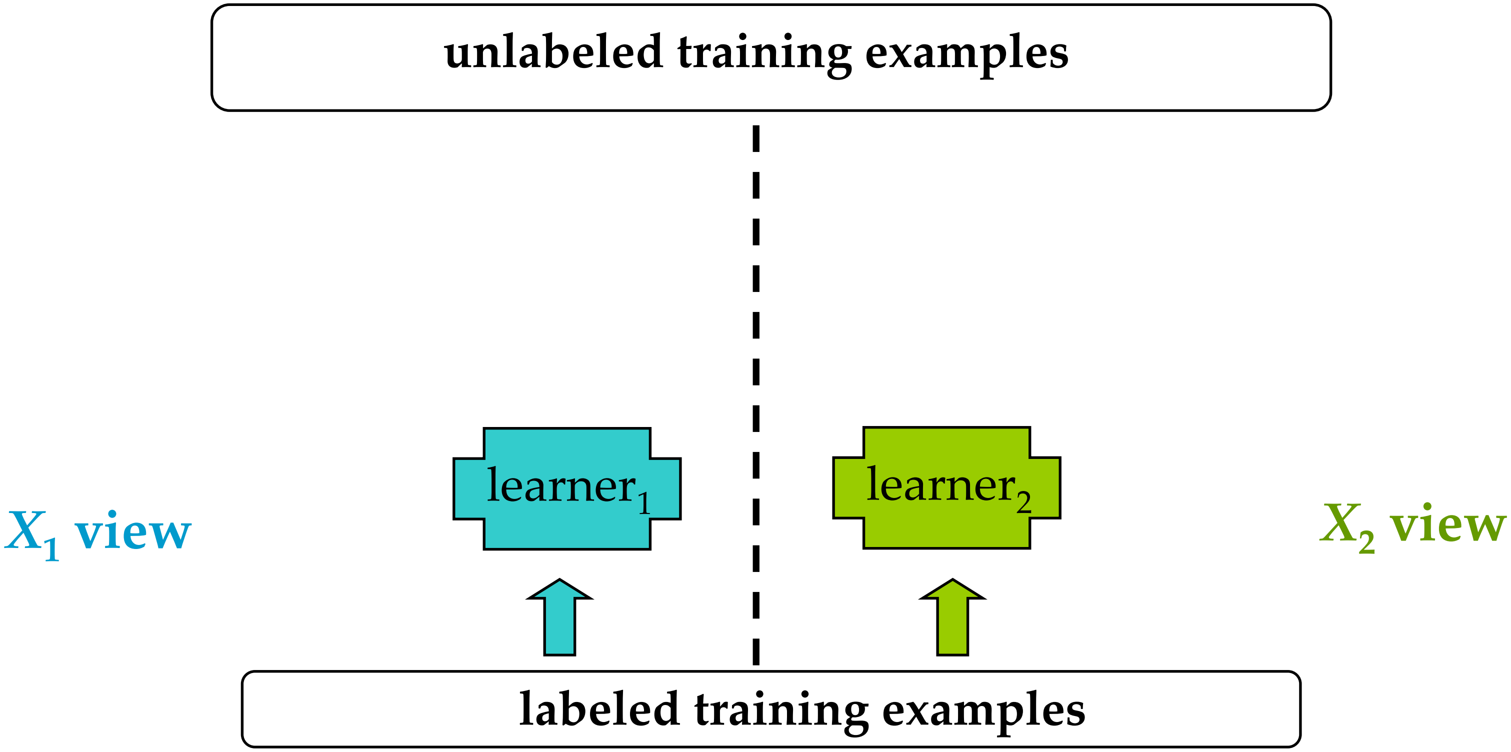
A previous survey showed about 83 percent of the general population believes in God.

But while medicine is science-based, doctors differ from scientists who work primarily in a laboratory setting, and their direct contact with patients in life-and-death situations may explain the differing views, Curlin said.

The study is based on responses to questionnaires mailed in 2003. It is to appear in an upcoming issue of the Journal of General Internal Medicine and was released online to subscribers earlier this month.

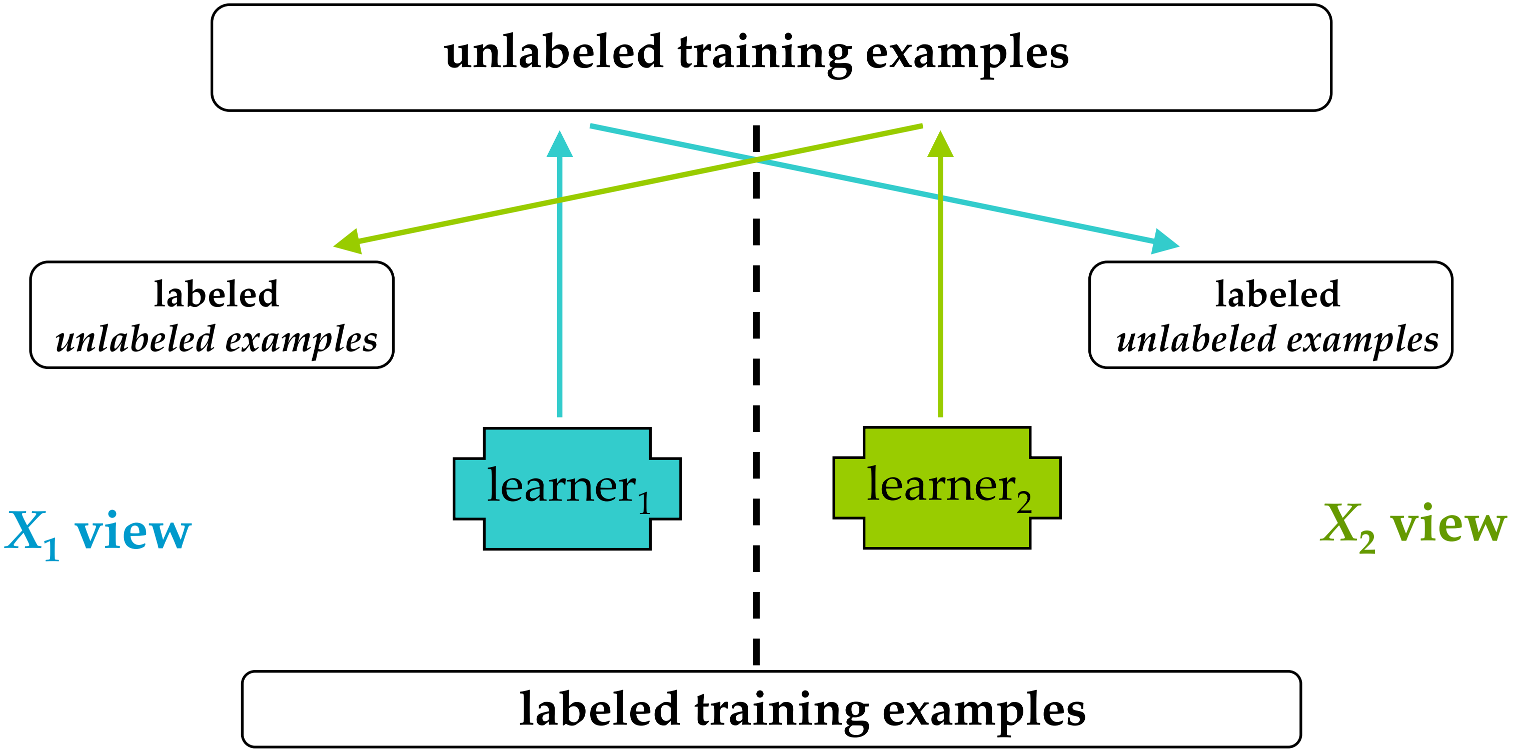
[A. Blum & T. Mitchell, COLT'98]

Co-training (con't)



[A. Blum & T. Mitchell, COLT'98]

Co-training (con't)



unlabeled training examples

labeled
unlabeled examples

labeled
unlabeled examples

learner_1

learner_2

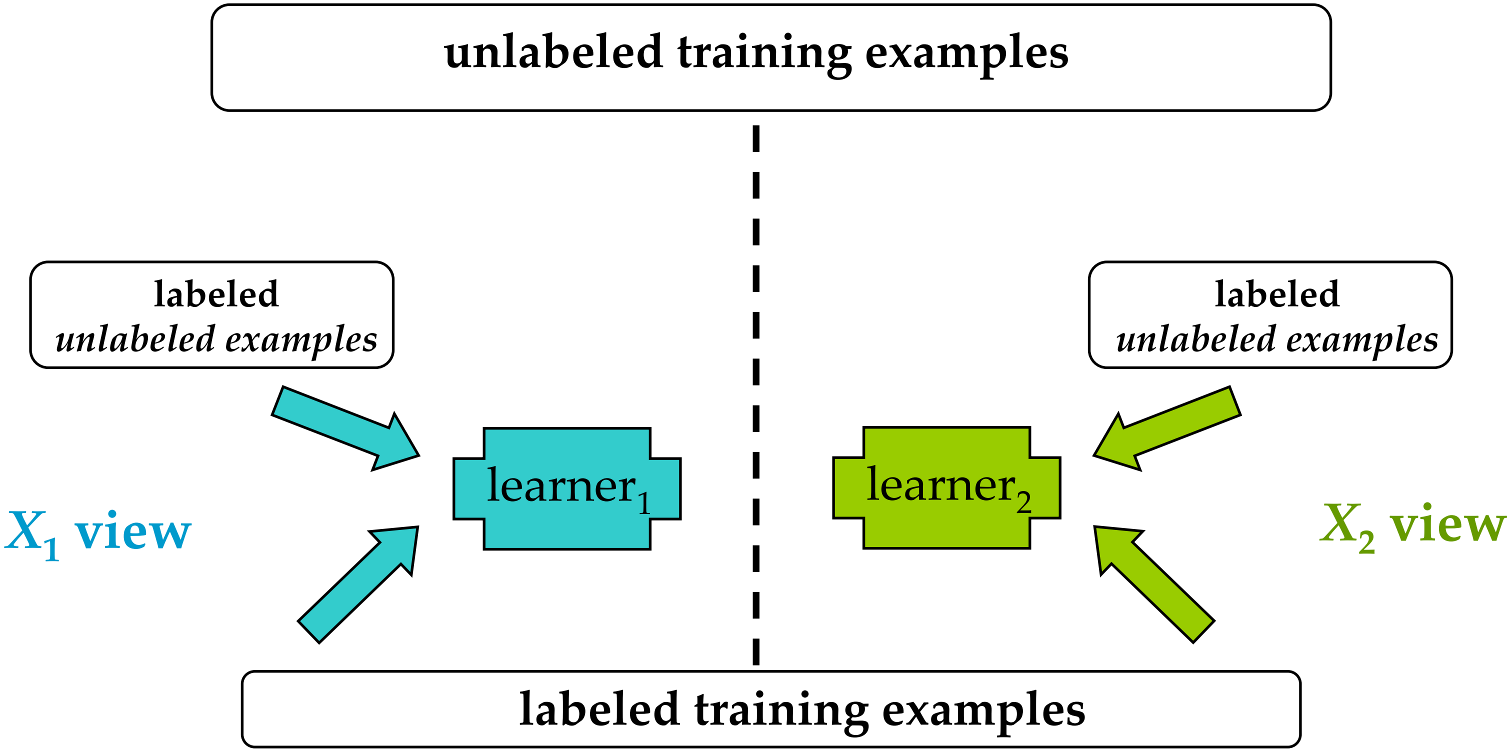
X_1 view

X_2 view

labeled training examples

[A. Blum & T. Mitchell, COLT'98]

Co-training (con't)



unlabeled training examples

labeled
unlabeled examples

labeled
unlabeled examples

X_1 view

learner$_1$

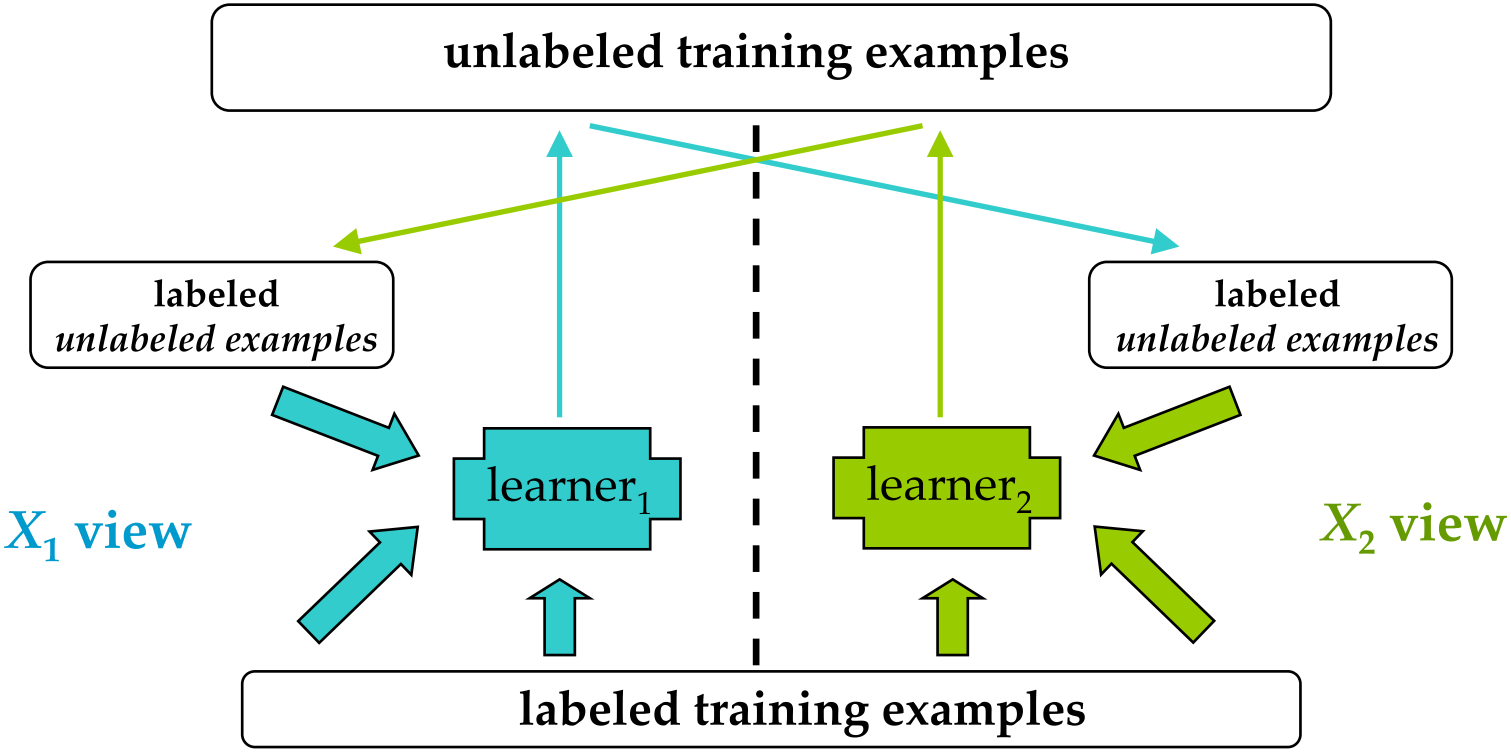
learner$_2$

X_2 view

labeled training examples

[A. Blum & T. Mitchell, COLT'98]

Co-training (con't)



[A. Blum & T. Mitchell, COLT'98]

Theoretical results

[A. Blum & T. Mitchell, COLT'98] - Given a conditional independence assumption on the distribution *D*, if the target class is learnable from random classification noise in the standard PAC model, then any initial weak predictor can be boosted to arbitrarily high accuracy by co-training

[S. Dasgupta et al., NIPS'01] – When the requirement of sufficient and redundant views is met, the co-trained classifiers could make few generalization errors by maximizing their agreement over the unlabeled data

[M.-F. Balcan et al., NIPS'04] - Given appropriately strong PAC-learners on each view, a weaker "expansion" assumption on the underlying data distribution is sufficient for iterative co-training to succeed

Applications

Although the requirement of sufficient and redundant views is quite difficult to meet, co-training has already been used in many domains, e.g.,

- Statistical parsing [A. Sarkar, NAACL01; M. Steedman et al., EACL03; R. Hwa et al., ICML03w]
- Noun phrase identification [D. Pierce & C. Cardie, EMNLP01]
- Image retrieval [Z.-H. Zhou et al., ECML'04, TOIS06]
- … …

Single-view variant

[S. Goldman & Y. Zhou, ICML'00] used two different supervised learning algorithms whose hypothesis partitions the example space into a set of equivalent classes

e.g. for a decision tree each leaf defines an equivalent class

Actually they used the ID3 decision tree and HOODG decision tree

Two key issues:

- How to combine the two classifiers?

  Using 10-fold CV to estimate the predictive confidence of the two classifiers and the involved equivalent classes

- How to choose unlabeled instance to label?

  Using 10-fold CV to estimate the labeling confidence

Weakness: Time-consuming 10-fold CV is used for many times in every round of the co-training process

Tri-training

The intuition:

If three classifiers are involved, maybe it is not necessary to measure the labeling confidence explicitly

- ➢ if two classifiers agree, then label for the other classifier
- ➢ the prediction can be made by voting these three classifiers

Additional benefit:

- Ensemble learning can be utilized to improve the generalization

[Z.-H. Zhou & M. Li, TKDE05]

Tri-training (con't)

A problem:

"Majority teach minority" may be wrong in some cases

- If the prediction of h_2 and h_3 on x is correct,
  then h_1 will receive a valid new example for further training

- Otherwise,
  h_1 will get an example with noisy label

  however, even in the worse case, the increase in the classification noise rate can be compensated if the amount of newly labeled examples is sufficient, under certain conditions

[Z.-H. Zhou & M. Li, TKDE05]

Tri-training (con't)

According to [D. Angluin & P. Laird, MLJ88], if a sequence σ of m samples is drawn, where the sample size m satisfies

$$m \geq \frac{2}{\epsilon^2 (1 - 2\eta)^2} \ln \left(\frac{2N}{\delta}\right)$$

ε : the hypothesis worst-case classification error rate
η (< 0.5): an upper bound on the classification noise rate
N: the number of hypothesis
δ: the confidence

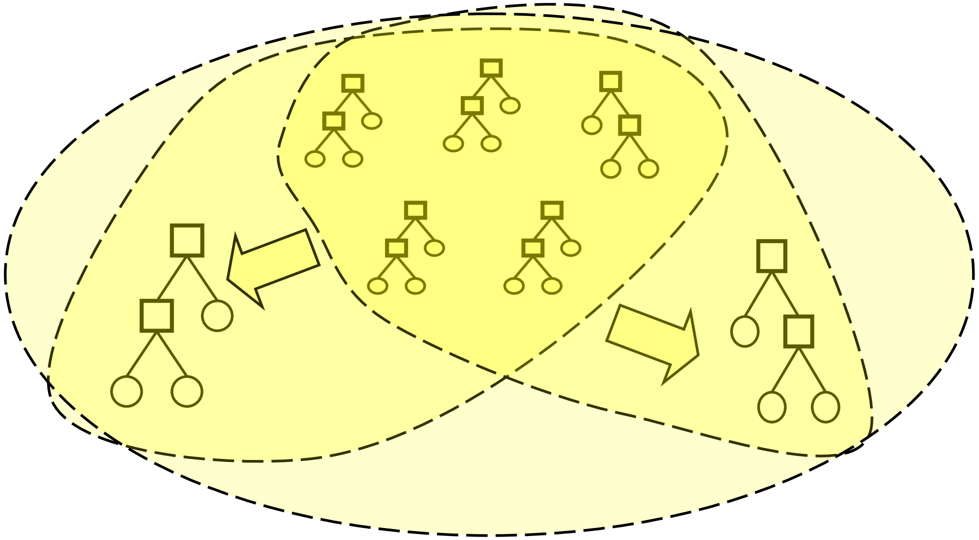
then a hypothesis H_i that minimizes disagreement with σ will have the PAC property: $\Pr\left[d(H_i, H^*) \geq \epsilon\right] \leq \delta$

From this we derived the tri-training criterion:

$$0 < \frac{\breve{e}_1^t}{\breve{e}_1^{t-1}} < \frac{|L^{t-1}|}{|L^t|} < 1$$

[Z.-H. Zhou & M. Li, TKDE05]

Co-Forest



Error of base classifier:

⇨ **Reduce**

Diversity among base classifier:

⇨ **Reduce**

Maintaining the *Diversity* during learning

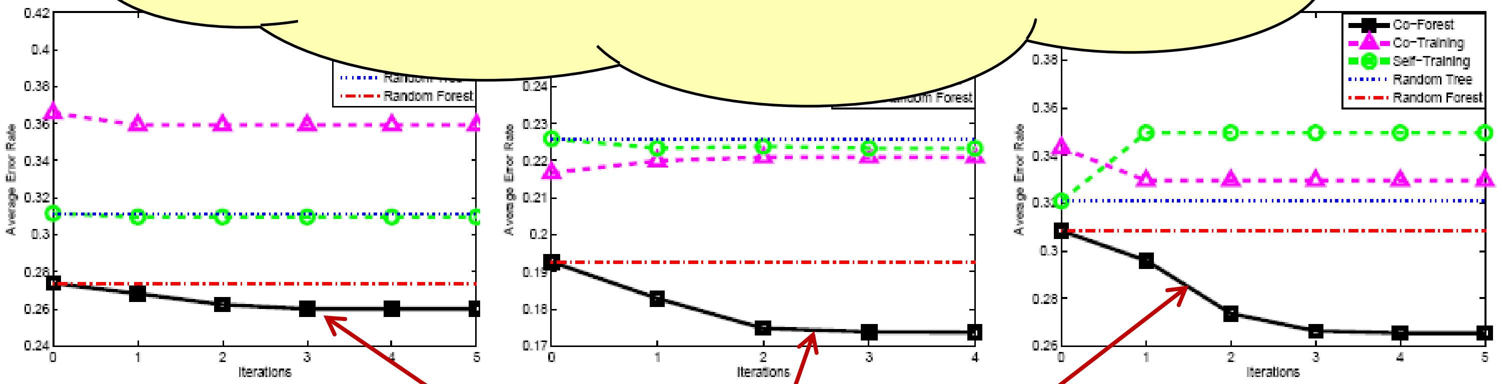
- Injecting **Randomness** (RF)
- Selecting unlabeled from an **unlabeled example pool**

[M. Li & Z.-H. Zhou, TSMCA07]

Co-Forest (con't)

Data set	RTree	Forest	SVM	AdaBoost	Self-Training			Co-Training			Co-Forest		
					initial	final	improv.	initial	final	improv.	initial	final	improv.
bupa	.396	.395	.420	.387	.396	.424	-7.1%*	.427	.443	-3.6%	.395	.384	**2.9%***
colic	.272	.208	.233	.230	.272	.278	-2.3%	[illegible]	.285	[illegible]	.208	.178	**14.5%***
diabetes	.321	.278	.261	.263								.261	**6.2%***
hepatitis	.231											[illegible]	**11.5%***
hypothyroid	.023												**6.6%**
ionosphere	[illegible]												[illegible]
kr-vs-kp													[illegible]
sonar													[illegible]
vote													[illegible]
wpbc													[illegible]
avg.													**14.2%**

Co-Forest gains better generalization ability by utilizing unlabeled data and utilizing ensemble

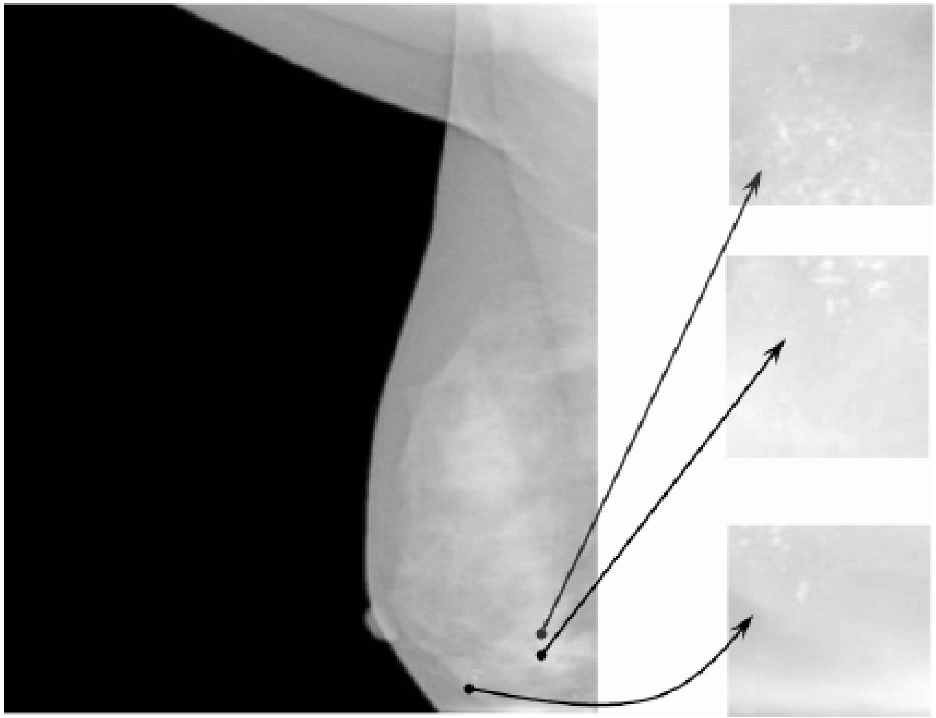


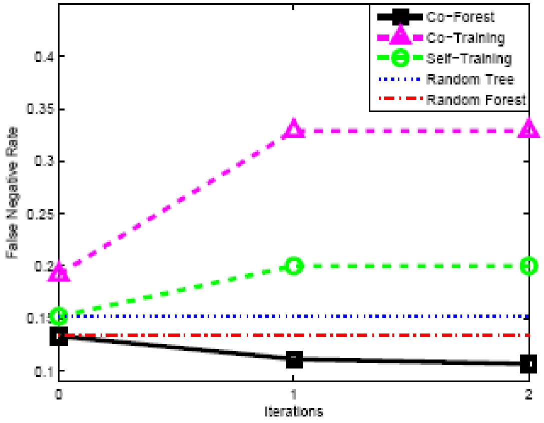
(a) *diabetes* (b) *hepatitis* (c) *wpbc*

Co-Forest

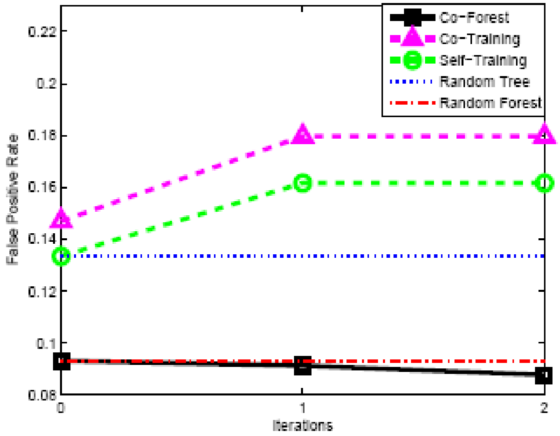
Co-Forest (con't)

Application to Microcalcification Detection





(a) *False Negative Rate*



(b) *False Positive Rate*

Co-Forest can help to reduce the false-negative rate while maintaining the false-positive rate by utilizing undiagnosed samples

[M. Li & Z.-H. Zhou, TSMCA07]

Other SSL ensemble methods

Semi-supervised Boosting methods:

- ✓ SS MarginBoost [F. d'Alché-Buc et al., NIPS'01]
- ✓ ASSEMBLE.AdaBoost [K. Bennett et al., KDD'02]

  Winner of the NIPS'01 Unlabeled Data Competition
- ✓ SemiBoost [P.K. Mallapragada et al., TPAMI in press]
- ✓ Multi-class SSBoost [H. Valizadegan et al., ECML'08]

Comparing with the huge amount of literatures on semi-supervised learning and ensemble learning, the literatures on SSL ensemble methods are too few

Problem

"Despite the theoretical and practical relevance of semi-supervised classification, the proposed approaches so far dealt with only single classifiers, and, in particular, no work was clearly devoted to this topic within the MCS literature"

Fabio Roli, MCS'05 Keynote

- □ SSL: Using unlabeled data is sufficient, why bother multiple learners?
- □ Ensemble: Using MCS is sufficient, why need unlabeled data?

Outline

- ➢ Ensemble Learning
- ➢ Semi-Supervised Learning
- ➢ **Classifier Combination vs. Unlabeled Data**
  - ✓Is classifier combination helpful to SSL ?
  - ✓Are unlabeled data helpful to ensemble ?
- ➢ Conclusion

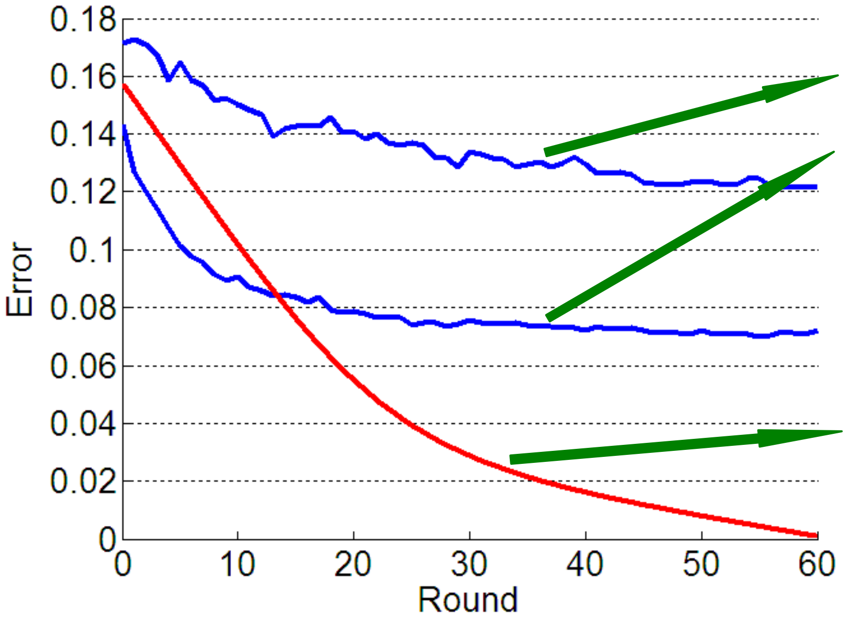
Single or combination?

In many SSL studies, it was shown that very strong classifiers can be attained by using unlabeled data

e.g., [A. Blum & T. Mitchell, COLT'98] - Given a conditional independence assumption on the distribution *D*, if the target class is learnable from random classification noise in the standard PAC model, then **any initial weak predictor can be boosted to arbitrarily high accuracy by co-training**

So, a single classifier seems enough

However, in empirical studies …



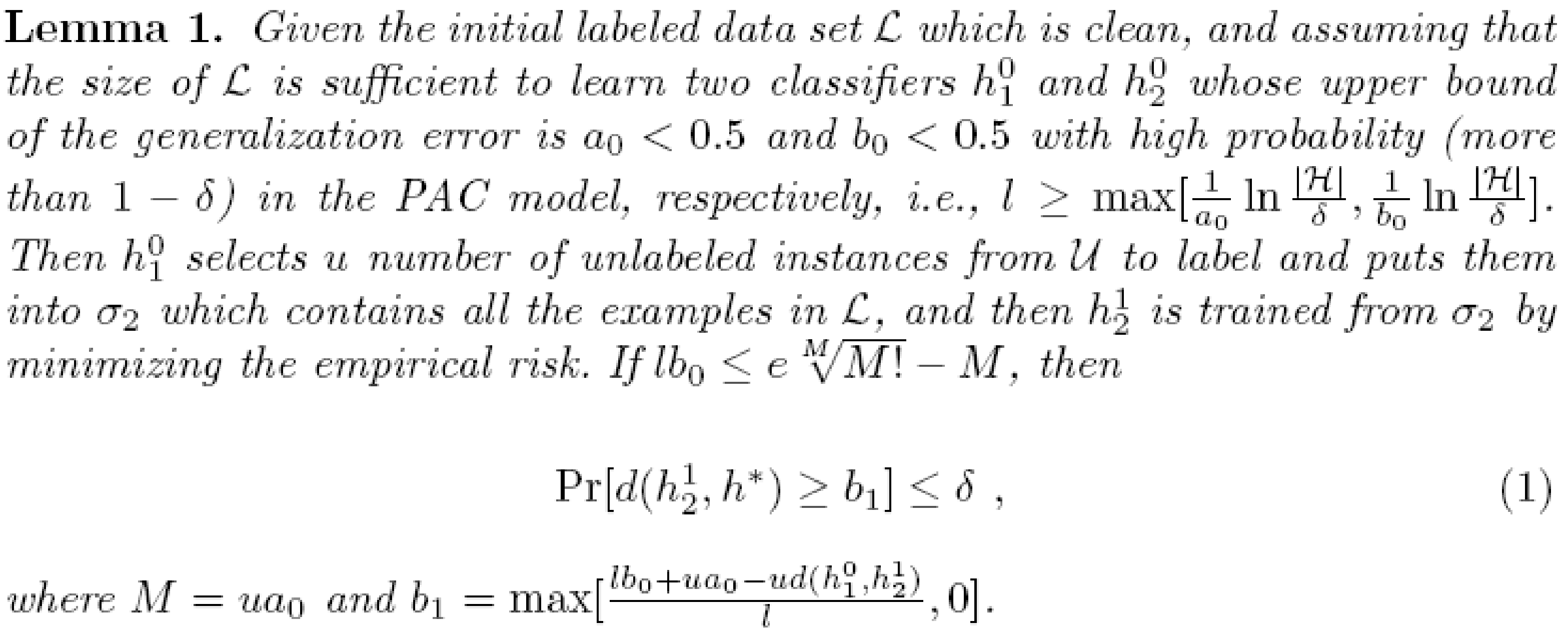
Performance of Co-training

Performances of the learners observed in experiments : the performances could not be improved further after a number of rounds

why?

Previous theoretical studies indicated that the performances could always be improved

Condition for co-training to work



Lemma 1. *Given the initial labeled data set $\mathcal{L}$ which is clean, and assuming that the size of $\mathcal{L}$ is sufficient to learn two classifiers h_1^0 and h_2^0 whose upper bound of the generalization error is $a_0 < 0.5$ and $b_0 < 0.5$ with high probability (more than $1 - \delta$) in the PAC model, respectively, i.e., $l \geq \max[\frac{1}{a_0} \ln \frac{|\mathcal{H}|}{\delta}, \frac{1}{b_0} \ln \frac{|\mathcal{H}|}{\delta}]$. Then h_1^0 selects u number of unlabeled instances from $\mathcal{U}$ to label and puts them into σ_2 which contains all the examples in $\mathcal{L}$, and then h_2^1 is trained from σ_2 by minimizing the empirical risk. If $lb_0 \leq e \sqrt[M]{M!} - M$, then*

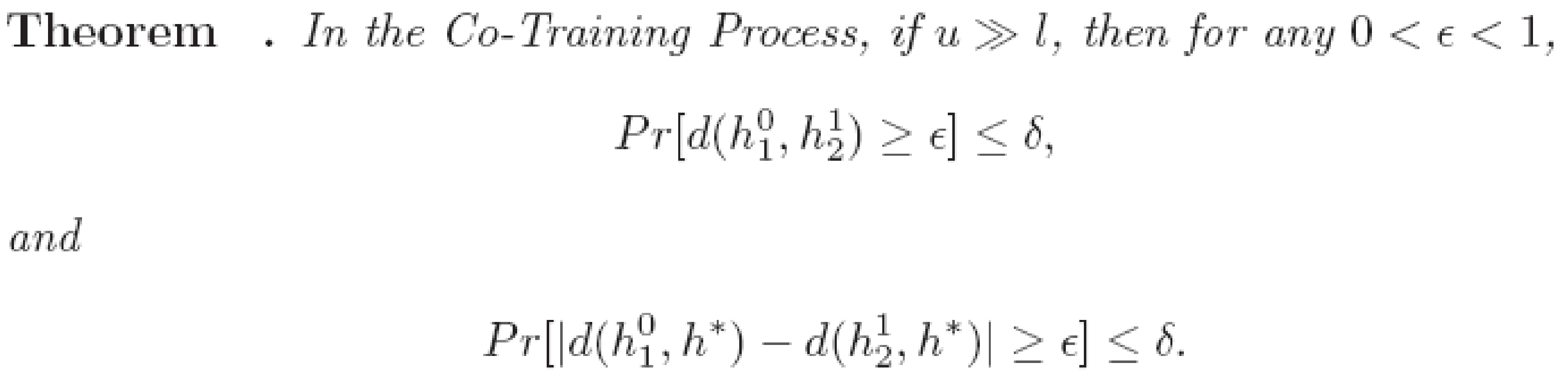
$$\Pr[d(h_2^1, h^*) \geq b_1] \leq \delta \ , \qquad (1)$$

where $M = ua_0$ *and* $b_1 = \max[\frac{lb_0 + ua_0 - ud(h_1^0, h_2^1)}{l}, 0]$.

Roughly speaking, the key requirement of co-training is that the initial learners should have large difference; it is not important that whether the difference is achieved by exploiting two views or not

[W. Wang & Z.-H. Zhou, ECML'07]

Is the theoretical/empirical gap occasional?



Theorem . *In the Co-Training Process, if $u \gg l$, then for any $0 < \epsilon < 1$,*

$$Pr[d(h_1^0, h_2^1) \geq \epsilon] \leq \delta,$$

and

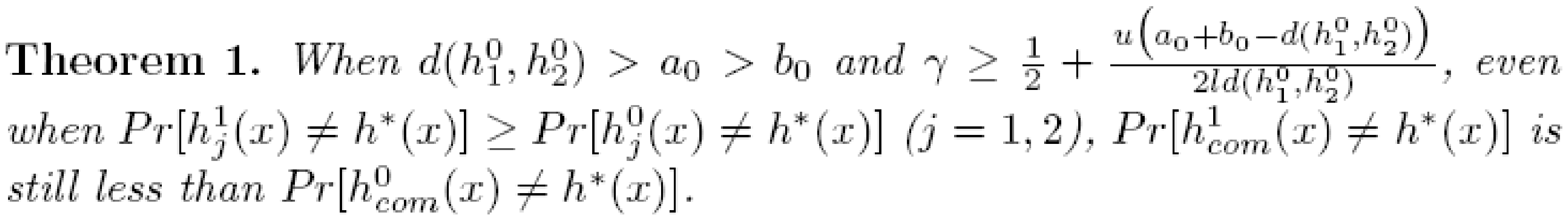
$$Pr[|d(h_1^0, h^*) - d(h_2^1, h^*)| \geq \epsilon] \leq \delta.$$

Roughly speaking, as the co-training process continues, the learners will become more and more similar, and therefore it is a “must”-phenomenon that co-training could not improve the performance further after a number of iterations

[W. Wang & Z.-H. Zhou, ECML’07]

Will classifier combination help?

"Later Stop"

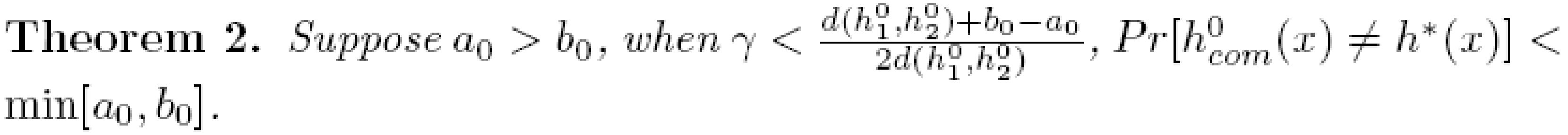


Theorem 1. *When* $d(h_1^0, h_2^0) > a_0 > b_0$ *and* $\gamma \geq \frac{1}{2} + \frac{u\left(a_0+b_0-d(h_1^0,h_2^0)\right)}{2ld(h_1^0,h_2^0)}$, *even when* $Pr[h_j^1(x) \neq h^*(x)] \geq Pr[h_j^0(x) \neq h^*(x)]$ $(j = 1, 2)$, $Pr[h_{com}^1(x) \neq h^*(x)]$ *is still less than* $Pr[h_{com}^0(x) \neq h^*(x)]$.

Roughly speaking, even when the individual learners could not improve the performance any more, classifier combination is still possible to improve generalization further by using more unlabeled data

To appear in a longer version of [W. Wang & Z.-H. Zhou, ECML'07]

"Earlier Success"



Theorem 2. *Suppose* $a_0 > b_0$, *when* $\gamma < \frac{d(h_1^0,h_2^0)+b_0-a_0}{2d(h_1^0,h_2^0)}$, $Pr[h^0_{com}(x) \neq h^*(x)] < \min[a_0, b_0]$.

Roughly speaking, the classifier combination is possible to reach a good performance earlier than the individual classifiers

To appear in a longer version of [W. Wang & Z.-H. Zhou, ECML'07]

Outline

- ➢ Ensemble Learning
- ➢ Semi-Supervised Learning
- ➢ **Classifier Combination vs. Unlabeled Data**
  - ✓Is classifier combination helpful to SSL ?
  - ✓Are unlabeled data helpful to ensemble ?
- ➢ Conclusion

First reason

When there are very few labeled training examples, ensemble could not work

SSL may be able to enable ensemble learning in such situation

At least how many labeled examples are needed for SSL ?

OLTV (One Labeled example and Two Views)

We show that when there are two sufficient views, SSL with a single labeled example is possible

$\mathcal{X}$ and $\mathcal{Y}$ – two views

$(\langle \boldsymbol{x}, \boldsymbol{y} \rangle, c)$ - a labeled example

where $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$ are the two portions of the example,
$c \in \{0, 1\}$ is the label

Assuming there exist functions $f_{\mathcal{X}}$ over $\mathcal{X}$ and $f_{\mathcal{Y}}$ over $\mathcal{Y}$, satisfying $f_{\mathcal{X}}(\boldsymbol{x}) = f_{\mathcal{Y}}(\boldsymbol{y}) = c$

which means that both are sufficient views

The Task:

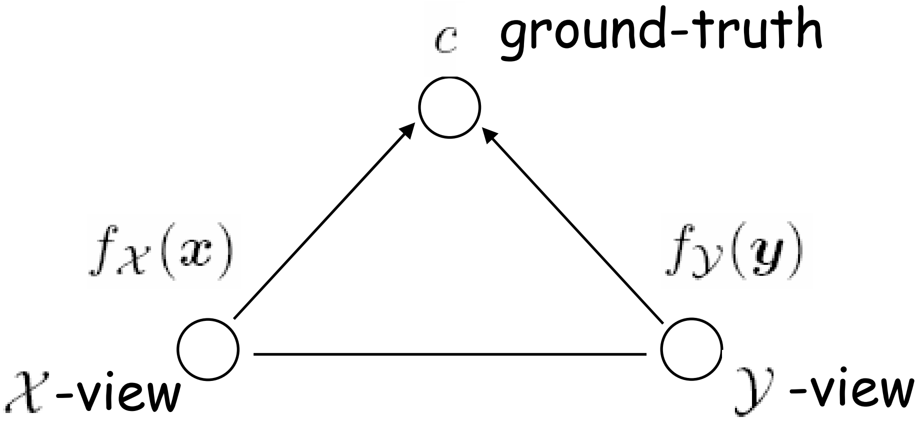
Given $(\langle \boldsymbol{x}_0, \boldsymbol{y}_0 \rangle, 1)$ and unlabeled examples $\mathcal{U} = \{(\langle \boldsymbol{x}_i, \boldsymbol{y}_i \rangle, c_i)\}$ ($i = 1, 2, \ldots, l$-1; c_i is unknown), to train a classifier

[Z.-H. Zhou et al., AAAI'07]

OLTV (con't)

For a sufficient view there should exist at least one projection which is correlated strongly with the ground-truth

If two sufficient views are conditionally independent given the class label, the most strongly correlated pair of projections should be in accordance with the ground-truth


c
ground-truth
$f_{\mathcal{X}}(\boldsymbol{x})$
$f_{\mathcal{Y}}(\boldsymbol{y})$
$\mathcal{X}$-view
$\mathcal{Y}$ -view

CCA (canonical correlation analysis) [Hotelling, Biometrika1936] can be used

[Z.-H. Zhou et al., AAAI'07]

OLTV (con't)

A number of correlated pairs of projections will be identified. The strength of the correlation can be measured by λ

m - the number of pairs of correlated projections that have been identified

$sim_{i,j}$ - the similarity between $<x_i, y_i>$ and $<x_0, y_0>$ in the j-th projection

$sim_{i,j}$ can be defined in many ways, such as:

$$sim_{i,j} = \exp\left(-d^2\left(P_j\left(x_i\right), P_j\left(x_0\right)\right)\right) + \exp\left(-d^2\left(P_j\left(y_i\right), P_j\left(y_0\right)\right)\right)$$

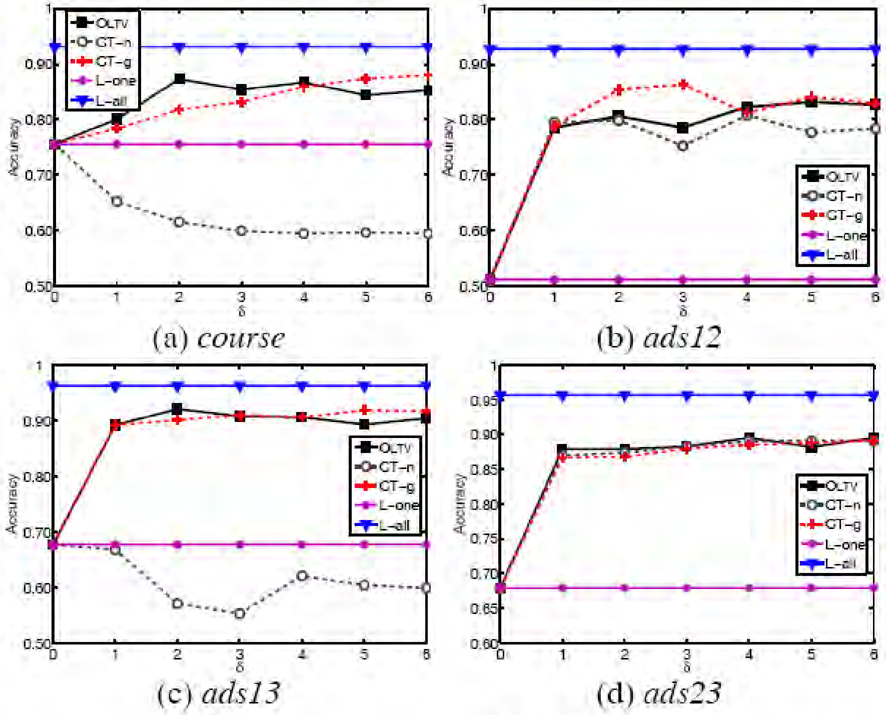
Then, the confidence of $<x_i, y_i>$ being a positive instance can be estimated:

$$\rho_i = \sum\nolimits_{j=1}^{m} \lambda_j sim_{i,j}$$

Thus, several unlabeled instances with the highest and lowest ρ values can be picked out respectively to be used as extra positive and negative instances

[Z.-H. Zhou et al., AAAI'07]

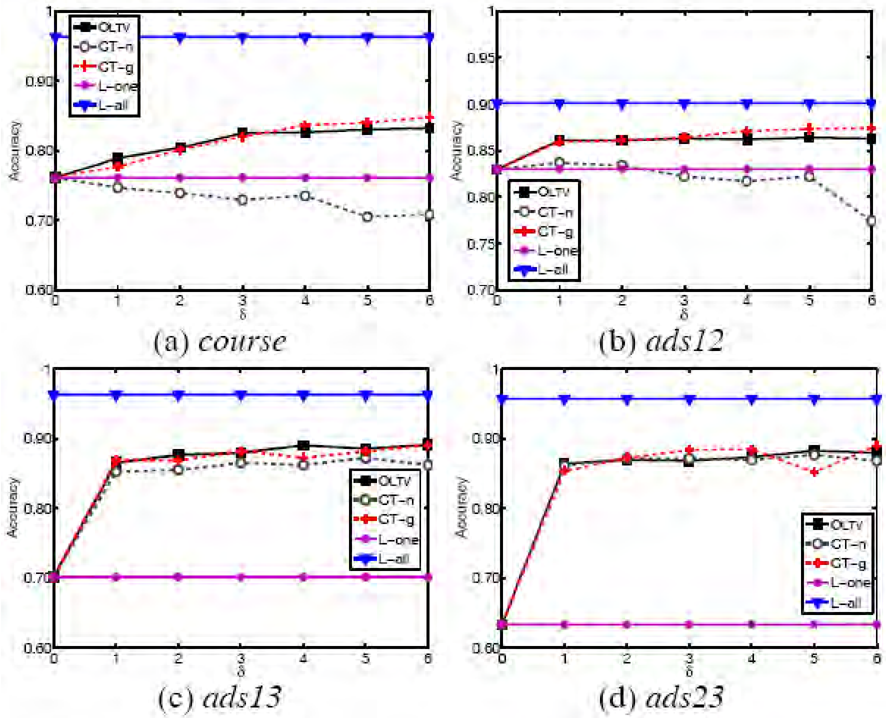
OLTV (con't)



(a) *course* (b) *ads12*

(c) *ads13* (d) *ads23*

Figure 1: Predictive accuracy with Naïve Bayes classifiers



(a) *course* (b) *ads12*

(c) *ads13* (d) *ads23*

Figure 2: Predictive accuracy with J48 decision trees

Figures reprinted from [Z.-H. Zhou et al., AAAI'07]

Second reason (possibly more important)

Diversity among the base learners is (possibly) the key of ensembles

Unlabeled data can be exploited for diversity-augment

A preliminary method

Basic idea:

In addition to maximize accuracy and diversity on labeled data, maximizing diversity on unlabeled data

Labeled training set : $\mathcal{L} = \{(x_1, y_1), \cdots, (x_l, y_l)\}$

Unlabeled training set : $\mathcal{U} = \{u_1, \cdots, u_n\}$

Unlabeled data set derived from $\mathcal{L}$: $\tilde{\mathcal{L}} = \{x_1, \cdots, x_l\}$

Assume the ensemble $\mathcal{E}$ consists of *m* linear classifiers $\{w_1 \cdots, w_m\}$ where w_k is weight vector of the *k*-th classifier

$\mathbf{W} = [w_1, \cdots, w_m]$ is the matrix formed by concatenating $\boldsymbol{w}_k$'s

A preliminary method (con't)

Generate the ensemble by minimizing the loss function:

$$V(\mathcal{L},\mathcal{U},\boldsymbol{W}) = \frac{1}{2}\sum_{k=1}^{m}||\boldsymbol{w}_k||_2^2 + C_1 \cdot \boxed{V_{acc}(\mathcal{L},\boldsymbol{W})} + C_2 \cdot V_{div}(\mathcal{D},\boldsymbol{W})$$

loss on accuracy

$$V_{acc}(\mathcal{L},\boldsymbol{W}) = \sum_{k=1}^{m}\sum_{i=1}^{l} loss(\boldsymbol{w}_k,\boldsymbol{x}_i,y_i)$$

$$loss(\boldsymbol{w}_k,\boldsymbol{x}_i,y_i) = \begin{cases} 0 & \text{if } y_i\langle\boldsymbol{w}_k,\boldsymbol{x}_i\rangle \geq 1 \\ (1 - y_i\langle\boldsymbol{w}_k,\boldsymbol{x}_i\rangle)^2 & \text{if } y_i\langle\boldsymbol{w}_k,\boldsymbol{x}_i\rangle < 1 \end{cases}$$

A preliminary method (con't)

Generate the ensemble by minimizing the loss function:

$$V(\mathcal{L},\mathcal{U},\boldsymbol{W}) = \frac{1}{2}\sum_{k=1}^{m}||\boldsymbol{w}_k||_2^2 + C_1 \cdot V_{acc}(\mathcal{L},\boldsymbol{W}) + C_2 \cdot \boxed{V_{div}(\mathcal{D},\boldsymbol{W})}$$

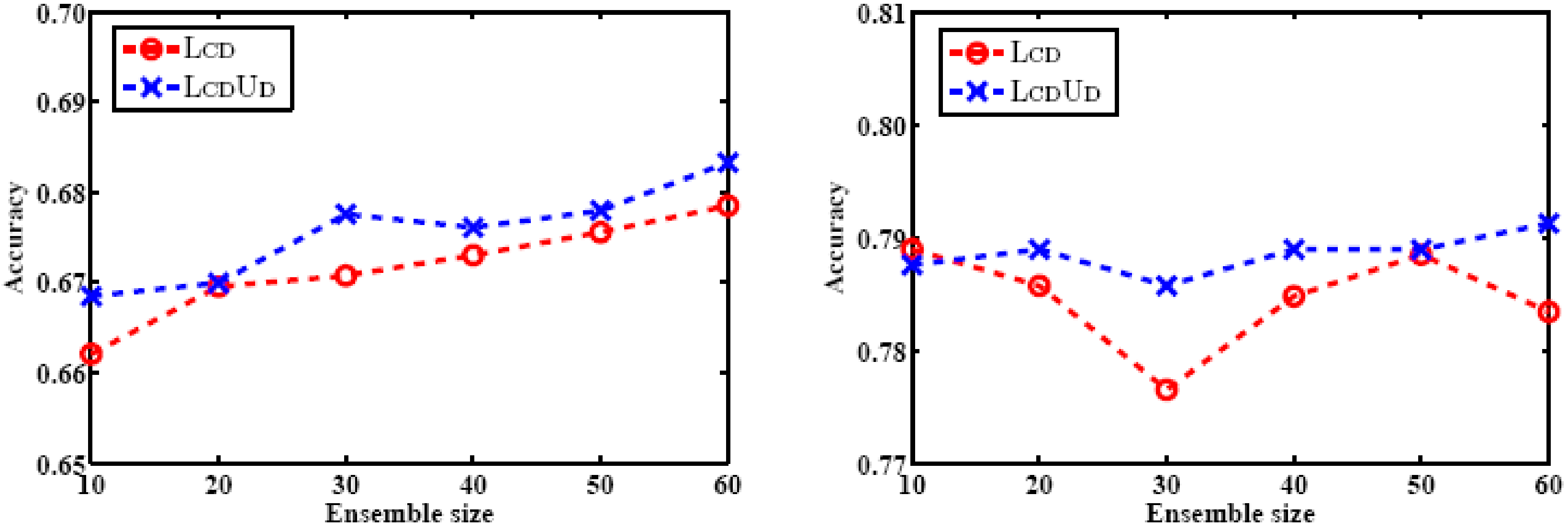
loss on diversity

$$V_{div}(\mathcal{D},\boldsymbol{W}) = \sum_{p=1}^{m-1}\sum_{q=p+1}^{m} d(\boldsymbol{w}_p,\boldsymbol{w}_q,\mathcal{D})$$

$$d(\boldsymbol{w}_p,\boldsymbol{w}_q,\mathcal{D}) = \begin{cases} 0 & \text{if } \mathcal{D}=\emptyset \\ \frac{\sum_{\boldsymbol{x}\in\mathcal{D}} \mathbf{sign}(\langle\boldsymbol{w}_p,\boldsymbol{x}\rangle)\cdot\mathbf{sign}(\langle\boldsymbol{w}_q,\boldsymbol{x}\rangle)}{|\mathcal{D}|} & \text{if } \mathcal{D}\neq\emptyset \end{cases}$$

We study two cases: L_{CD} $(\mathcal{D}=\tilde{\mathcal{L}})$ and $\mathrm{L}_{\mathrm{CD}}\mathrm{U}_{\mathrm{D}}$ $(\mathcal{D}=\tilde{\mathcal{L}}\bigcup\mathcal{U})$

Preliminary results



(a) g241n ($N = 1500,\ d = 241$)

(b) vehicle ($N = 435,\ d = 26$)

Fig. 1. Comparing the performance of LCD and LCDUD. N is the number of instances; d is the dimensionality.

Conclusion

Ensemble learning and Semi-supervised learning are mutually beneficial

□ Classifier Combination is helpful to SSL:

- Later Stop
- Earlier Success

□ Unlabeled Data is helpful to Ensemble:

- Enable ensemble with very few labeled data
- Diversity augment

Promising Future

Ensemble -> Strong Classifier

SSL -> Strong Classifier

Ensemble and SSL -> Strong^2 Classifier

Thanks!