

Combining Labeled and Unlabeled Data with Co-Training

Avrim Blum, Tom Mitchell

Proceedings of the Workshop on Computational Learning Theory, Morgan
Kaufmann Publishers, 1998

Saif Mohammad

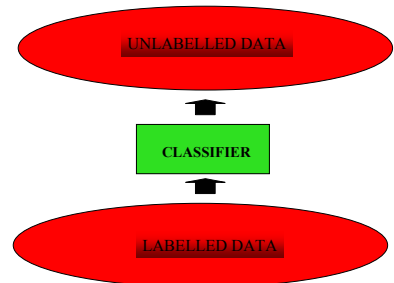
Lecturer: Rich Maclin

INTRODUCTION

Supervised Learning

- Learn classifier from annotated training data
- Apply classifier on unseen test data
- Larger the training set
 - More accurate is the classifier
- Consider Word Sense Disambiguation
 - Annotated / labeled data sample
Jack withdrew money from the [bank/financial_institution](#)
 - Raw data / unlabeled data sample
Banks have since long attracted robbers

Classical Approach



The Bottleneck

- Human annotation
 - Expensive
 - Time intensive
- Necessary for every domain / problem for which supervised machine learning applied
- Unsupervised learning
 - A possible solution
 - Low accuracies so far

Solution

Combining Labeled and Unlabeled Data
with Co-Training

- A **small** set of labelled data
- A large set of unlabelled data
- High accuracy in classification achieved using...

Co-Training

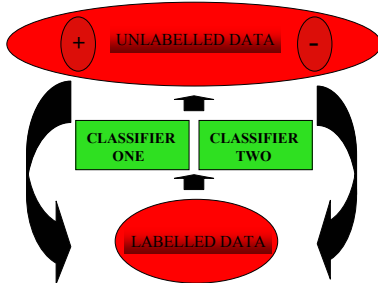
Outline

- Co-Training ←
- Examples
- PAC framework
- Experiment
- Conclusion

Informal Description

- Anoop Sarkar on Co-training – Applying Co-training Methods to Statistical Parsing
 - Pick (two) or more views of a classification problem
 - Build separate models for each view
 - Train each model on a small set of labeled data – initial training set
 - Classify unlabelled data
 - Pick examples which each model independently labels with high confidence
 - Add to the training set

Co-training Approach



Co-training: Re-stated

- Identify two “kinds” of information of the examples
- Use initial small set of labeled examples to train weak predictors based on the two kinds
- Bootstrap from the weak predictors using unlabelled data

This type of bootstrapping is called [Co-training](#)

Outline

- Co-Training
- Examples ←
- PAC framework
- Experiments
- Conclusions and Open Questions

EXAMPLES

Web Page Classification

Given a set of web pages from say the Computer Science Department websites, identify the course web pages.

- Labeled data
 - 1051 CS web pages from four universities
 - Course home pages – positive examples
 - All other pages – negative examples

Web Page Classification

- Weak predictors
 - Bag of words appearing on the web page
 - Course, syllabus, midterm, final, exam, quiz and so on
 - Bag of words underlined in all links pointing to the web page
 - Examples of such links
 - **CS8751 course syllabus, The Advanced ML course**
 - Course, syllabus, CS8751, advanced and so on
- May use a naïve bayes classifier to do the classification

Word Sense Disambiguation

To identify the intended sense of a word from a list of possible senses, based on its context

- Yarowsky's pioneering work in co-training
- Each example/instance represented in two views
 - Unique Document ID
 - Context of the word – collocations

The industrial plant besides lake Superior is cause of concern.
The plant is the major source of pollution in the lake.

The industrial plant besides lake Superior is cause of concern.
The plant is the major source of pollution in the lake.

Assumptions

- One sense per discourse
 - Two instances of a word in the same document have same sense
 - Plant₁ and plant₂, lake₁ and lake₂
- One sense per collocation
 - 'industrial plant' restricts possible senses of plant to 1
 - The context of the word sufficient for disambiguation

Algorithm

- Identify examples of polysemous word
- For each sense of the word
 - Identify a small number of seed collocations - C_i
 - Industrial plant, flowering plant
- Identify examples E_i in unlabelled corpus which have collocations in C_i and tag them with sense i
- Identify new collocations in E_i
 - Add them to C_i

Second View...and thus co-training

- One sense per discourse
 - If several instances of a word in a discourse are of sense A
 - Remaining instances in discourse tagged A
 - Testing and training done on various articles and abstracts
 - Each constituting one discourse
- This provides a bridge to attain new contexts
 - Giving new collocations
 - Collocations which may not be found close to already found collocations

Outline

- Co-Training
- Examples
- PAC framework ←
- Experiments
- Conclusion

THE PAC FRAMEWORK

Probably Approximately Correct

- We would like to be able to characterize classes of target concepts that can be reliably learned from a reasonable number of randomly drawn training examples
- Insight into relative complexity of different learning problems
- Rate at which accuracy improves with additional training examples

PAC - learnable

- Consider concept class C defined over a set of instances X and a learner L using hypothesis space H
 - C is PAC-learnable by L using H if for all $c \in C$, distributions D over X , ϵ such that $0 < \epsilon < 1/2$, and δ such that $0 < \delta < 1/2$, learner L with probability at least $(1 - \delta)$ output a hypothesis $h \in H$ such that $\text{error}_D(h) \leq \epsilon$, in time that is polynomial in $1/\epsilon$, $1/\delta$ and $\text{size}(C)$
- PAC provides a mathematical framework for assessing various concepts especially provides a bound to the number of examples needed to learn a concept successfully with probability $(1 - \delta)$ with an error $< \epsilon$

Is Co-training Justifiable by PAC ?

- Consider the rote learning problem
 - classifier remembers all training examples
 - If new example seen before then classified
 - Else 'I don't know'
- Define instance space $X = X_1 \times X_2$
 - X_1 and X_2 correspond to two views of an example
 - Let $|X_1| = |X_2| = N$
 - Probability that $(m+1)^{\text{st}}$ example has not yet been seen

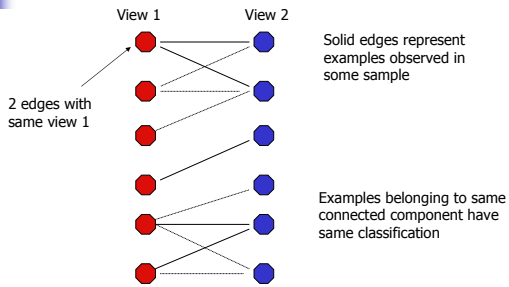
$$\sum_{x_1 \in X_1} P[x_1](1 - P[x_1])^m$$

Two views

- With co-training we have two views
- Rote learner confident when
 - Labeled example present in the connected component
 - If c_1, c_2, \dots, c_n are the connected components and have probabilities P_1, P_2, \dots, P_n
 - Probability that given m labeled examples, the label of $(m+1)^{\text{st}}$ example cannot be deduced:

$$\sum_{c_j \in G_D} P_j(1 - P_j)^m$$

Bipartite Graph




Using the Bipartite Graph

- Consider a bipartite graph G_s with one edge for every observed example
 - Let $|S|$ be number of edges in G_s , s_j is no. of components c_j
 - If a random subset m of the examples is labeled
 - m edges created
 - Probability that label of $(m+1)^{th}$ example cannot be deduced:

$$\sum_{c_j \in G_s} \frac{s_j [(|S| - s_j) \text{ choose } m]}{|S| \text{ choose } m + 1}$$
 - If $m \ll |S|$,

$$\sum_{c_j \in G_s} \frac{s_j}{|S|} \left(1 - \frac{s_j}{|S|}\right)^m$$

Outline

- Co-Training
- Examples
- PAC framework
- Experiment 
- Conclusion

EXPERIMENT

Web Page Classification

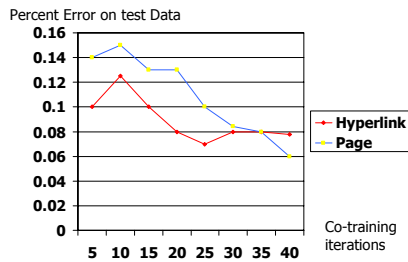
- 1051 web pages collected from CS department websites of four universities
- Task
 - To classify if web page is course syllabus page
- Two views
 - Bag of words in hyperlink
 - Bag of words in web page
- Classifier used
 - Naive Bayes

Results

Error Rates

	Page Based Classifier	Hyperlink based Classifier	Combined Classifier
Supervised Learning	12.9	12.4	11.1
Co-training	6.2	11.6	5.0

Do iterations help?



Outline

- Co-Training
- Examples
- PAC framework
- Experiments
- **Conclusion** ←

IN CONCLUSION

Is co-training worth it ?

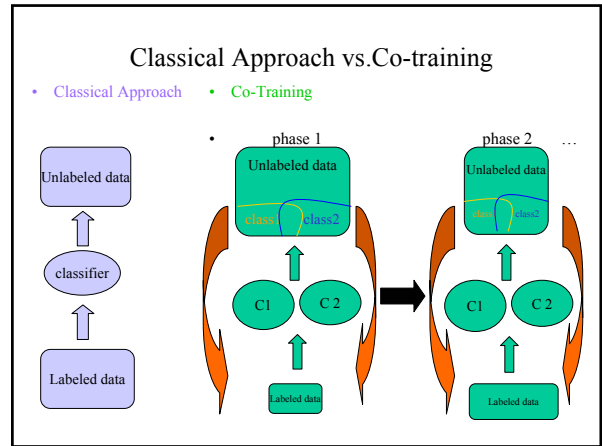
- Model in which unlabeled data can be used to augment labeled data, based on having two views.
- Preliminary experimental results : promising
- Theory – sound
- Large number of applications possible
 - Video n audio

References

- Avrim Blum, Tom Mitchell. Combining Labeled and Unlabeled Data with Co-Training. Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers, 1998
- Yarowsky, David. 1995. Unsupervised word sense disambiguation rivaling supervised methods. Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, pages 189–196.
- A. Sarkar. Applying Co-Training Methods to Statistical Parsing. In Proceedings of NAACL

Combining labeled and unlabeled Data with Co-Training

Paper by: Avrim Blum, Tom Mitchell
 Presented by: Saif Mohammand
 Commented by: Ruinan Lu



Key Theoretical Prerequisites of Applying Co-training

- Compatibility
- Mutual independence:
 - x_1 and x_2 are conditionally independent given the label.
 - Theorem 1: If C_2 is learnable in the PAC model with classification noise, and if the conditional independence assumption is satisfied, then (C_1, C_2) is learnable in the co-training model from unlabeled data only, given an initial weakly-useful predictor $h(x_1)$.

Explore Consistency Condition

- Consider:
 - use unlabeled examples to prune away “incompatible” target concepts.

Explore Independence in Practical Problems...

- When features can be naturally split...
 - Identify televised segments containing the US president:
 - X_1 : set of possible video images.
 - X_2 : set of possible audio images.
 - Let $X = X_1 \times X_2 \dots$
 - Perception learning tasks involving multiple sensors: a mobile robot that must learn to recognize open doorways based on a collection of:
 - Vision: X_1
 - Sonar: X_2
 - Laser range sensor: X_3
 - Let $X = X_1 \times X_2 \times X_3 \dots$

Combining labeled and unlabeled data with co-training

By Blum and Michelle

Presented by: Saif

Comments by: Sweta

Analysing the effectiveness

- Assumptions
 - Instance distribution is compatible
 - Conditional independence
- Read world data does not satisfy this
- How sensitive co-training to the correctness of this assumption?
- What if there is no natural split?

Experiments

- What makes co-training better?
 - Each added document are informative – under assumption of conditional independence
- Results on WebKB – course dataset

Algorithm	# labeled	# Unlabeled	Error
Naive Bayes	788	0	3.3%
Co-training	12	776	5.4%
EM	12	776	4.3%
Naive Bayes	12	0	13.0%

Why co-training error high

- Too easy task
- Feature not sufficiently independent
- Do not adequately benefit from existing independence

Experiments on News 2*2 dataset

- Four newsgroups dataset
- Join document of first two newsgroup as + ve
- Join document of second two as -ve

Algorithm	# labeled	# Unlabeled	Error
Naive Bayes	1066	0	3.9%
Co-training	6	1000	3.7%
EM	6	1000	8.9%
Naive Bayes	6	0	34.0%

References

1. Combining labeled and unlabeled data with co-training, Blum and Mitchell, COLT-98
2. Analyzing the effectiveness and applicability of co-training, Nigam and Ghani.