A Co-Regularization Approach to Semi-supervised Learning with Multiple Views

Vikas Sindhwani VIKASS@CS.UCHICAGO.EDU
Partha Niyogi NIYOGI@CS.UCHICAGO.EDU
Mikhail Belkin MISHA@CS.UCHICAGO.EDU
Department of Computer Science, University of Chicago, Chicago, IL 60637

Abstract

The Co-Training algorithm uses unlabeled examples in multiple views to bootstrap classifiers in each view, typically in a greedy manner, and operating under assumptions of view-independence and compatibility. In this paper, we propose a Co-Regularization framework where classifiers are learnt in each view through forms of multi-view regularization. We propose algorithms within this framework that are based on optimizing measures of agreement and smoothness over labeled and unlabeled examples. These algorithms naturally extend standard regularization methods like Support Vector Machines (SVM) and Regularized Least squares (RLS) for multi-view semi-supervised learning, and inherit their benefits and applicability to high-dimensional classification problems. An empirical investigation is presented that confirms the promise of this approach.

1. Introduction

A striking aspect of natural learning is the ability to integrate and process multi-modal sensory information with very little supervisory feedback. The scarcity of labeled examples, abundance of unlabeled data and presence of multiple representations are aspects of several applications of machine learning as well. An example is hypertext classification: Modern search engines can index more than a billion web-pages in a single web-crawl, but only a few can be hand-labeled and assembled into web directories. Each web-page has disparate descriptions: textual content, inbound and outbound hyperlinks, site and directory names, etc. Although traditional machine learning has focussed on two extremes of an information spectrum (supervised and unsupervised learning), a number of recent efforts have considered the middle-ground of semi-supervised learning, with or without a multi-view component (Belkin, Matveeva, & Niyogi, 2004; Belkin, Niyogi & Sindhwani, 2004; Sindhwani, Niyogi & Belkin, 2005; Joachims, 1999; Joachims, 2003; Blum & Mitchell, 1998; Brefeld & Scheffer; Chapelle & Zien, 2005; Zhou et al, 2004).

The Co-Training framework proposed in (Blum & Mitchell, 1998) has been among the first efforts that provided a widely successful algorithm with theoretical justifications. The framework employs two assumptions that allow unlabeled examples in multiple-views to be utilized effectively: (a) the assumption that the target functions in each view agree on labels of most examples (*compatibility* assumption) and (b) the assumption that the views are independent given the class label (*independence* assumption). The first assumption allows the complexity of the learning problem to be reduced by the constraint of searching over compatible functions; and the second assumption allows high performance to be achieved since it becomes unlikely for compatible classifiers trained on independent views to agree on an incorrect label. The co-training idea has become synonymous with a greedy agreement-maximization algorithm that is initialized by supervised classifiers in each view and then iteratively re-trained on boosted labeled sets, based on high-confidence predictions on the unlabeled examples. The original implementation in (Blum & Mitchell, 1998) runs this algorithm on naive-bayes classifiers defined in each view. For more on agreement maximization principles, see (Abney, 2002; Dasgupta, Littman & McAllester, 2001; Collins & Singer, 1999; Yarowsky, 1995).

In this paper, we present a *Co-Regularization* framework for multi-view semi-supervised learning. Our approach is based on implementing forms of multi-view

Appearing in *Proceedings of the Workshop on Learning with Multiple Views*, 22^{nd} ICML, Bonn, Germany, 2005.

regularization using unlabeled examples. We suggest a family of algorithms within this framework: The Co-Regularized Least Squares (Co-RLS) algorithm performs a joint regularization that attempts to minimize disagreement in a least squared sense; the Co-Regularized Laplacian SVM and Least Squares (Co-LapSVM, Co-LapRLS) algorithms utilize multi-view graph regularizers to enforce complementary and robust notions of smoothness in each view. The recently proposed Manifold Regularization techniques (Belkin, Niyogi & Sindhwani, 2004; Sindhwani, 2004; Sindhwani, Niyogi & Belkin, 2005) are employed for Co-LapSVM and Co-LapRLS. Learning is performed by effectively exploiting useful structures collectively revealed with multiple representations.

We highlight features of the proposed algorithms:

1. These algorithms arise from natural extensions of the classical framework of regularization in Reproducing Kernel Hilbert Spaces. The unlabeled data is incorporated via additional regularizers that are motivated from recognized principles of semi-supervised learning.

2. The algorithms are non-greedy, involve convex cost functions and can be easily implemented.

3. The influence of unlabeled data and multiple views can be controlled explicitly. In particular, single view semi-supervised learning and standard supervised algorithms are special cases of this framework.

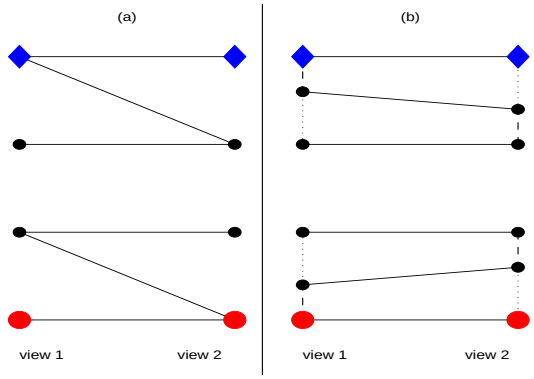
4. Experimental results demonstrate that the proposed methods out-perform standard co-training on synthetic and hypertext classification datasets.

In section 2, we setup the problem of semi-supervised learning in multiple views. In subsequent sections, we discuss the Co-Regularization framework, propose our algorithms and evaluate their empirical performance.

2. Multi-View Learning

In the multi-view semi-supervised learning setting, we have labeled examples $\{(x_i, y_i)\}_{i=1}^{l}$ and unlabeled examples $\{x_i\}_{l+1}^{l+u}$ where each example $x = (x^{(1)}, x^{(2)})$ is seen in two views with $x^{(1)} \in X^{(1)}$ and $x^{(2)} \in X^{(2)}$. The setup and the algorithms we discuss can also be generalized to more than two views. For the rest of this discussion, we consider binary classification problems where $y_i \in \{-1, 1\}$. The goal is to learn the function pair $f = (f^{(1)}, f^{(2)})$, where $f^{(1)} : X^{(1)} \mapsto \{-1, 1\}$ and $f^{(2)} : X^{(2)} \mapsto \{-1, 1\}$ are classifiers in the two views. In this paper, we will focus on how the availability of unlabeled examples and multiple views may be profitably leveraged for learning high-performance classifiers $f^{(1)}, f^{(2)}$ in each view.

Figure 1. Bipartite Graph Representation of multi-view learning. The small black circles are unlabeled examples.



How can unlabeled data and its multiple views help? In Figure 1(a), we reproduce the bipartite graph representation of the co-training setting, to initiate a discussion. The figure shows the two views of labeled and unlabeled examples, arranged as a bipartite graph. The left and right nodes in the graph are examples as seen in view 1 and view 2 respectively, with edges connecting the two views of an example. The unlabeled examples are shown as small black circles and the other examples are labeled. The class of compatible pairs of functions identically label two nodes in the same connected component of this graph. This may be interpreted as a requirement of smoothness over the graph for the pair $(f^{(1)}, f^{(2)})$. Thus, unlabeled examples provide empirical estimates of regularizers or measures of smoothness to enforce the right complexity for the pair $(f^{(1)}, f^{(2)})$.

In many applications, it is unrealistic for two examples to share a view exactly. A more realistic situation is depicted in Figure 1(b) where three types of edges are shown: (solid) edges connecting views of each example as in Figure 1(a); (dashed) edges connecting similar examples in each view; and (dotted) edges connecting examples in each view based on similarity in the other view. The similarity structure in one view induces a complementary notion of similarity in the other views with respect to which regularizers can be constructed using unlabeled data.

In the next section, we describe algorithms that arise from constructions of such regularizers.

3. Co-Regularization

The classical regularization framework (Poggio & Girosi, 1990; Schoelkopf & Smola, 2002; Vapnik, 1998) for supervised learning solves the following minimization problem :

$$f^* = \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{l} \sum_{i=1}^{l} V(x_i, y_i, f) + \gamma \|f\|_K^2 \qquad (1)$$

where $\mathcal{H}_K$ is an Reproducing Kernel Hilbert space (RKHS) of functions with kernel function K; $\{(x_i, y_i)\}_{i=1}^{l}$, is the labeled training set; and V is some loss function, such as squared loss for Regularized Least Squares (RLS) or the hinge loss function for Support Vector Machines (SVM). By the Representer theorem, the minimizer is a linear combination of kernel functions centered on the data:

$$f(x) = \sum_{i=1}^{l} \alpha_i K(x, x_i)$$

This real-valued function is thresholded and used for binary classification.

In the Co-regularization framework, we attempt to learn the pair $f = (f^{(1)}, f^{(2)})$ in a cross-product of two RKHS defined over the two views, i.e., $f^{(1)} \in \mathcal{H}_{K^{(1)}}$ and $f^{(2)} \in \mathcal{H}_{K^{(2)}}$. The key issue is imposing an appropriate notion of complexity on this pair so that a regularized solution effectively utilizes unlabeled data in the two views. We now describe some ideas.

Co-Regularized Least Squares

A natural idea is to attempt to learn the pair $f = (f^{(1)}, f^{(2)})$ so that each function correctly classifies the labeled examples, and the outputs of the pair agree over unlabeled examples. This suggests the following objective function:

$$(f^{(1)*}, f^{(2)*}) = \underset{\substack{f^{(1)} \in \mathcal{H}_{K^{(1)}} \\ f^{(2)} \in \mathcal{H}_{K^{(2)}}}}{\operatorname{argmin}} \sum_{i=1}^{l} \left[y_i - f^{(1)}(x_i^{(1)}) \right]^2 +$$

$$\mu \sum_{i=1}^{l} \left[y_i - f^{(2)}(x_i^{(2)}) \right]^2 + \gamma_1 \|f^{(1)}\|^2_{\mathcal{H}_{K^{(1)}}} +$$

$$\gamma_2 \|f^{(2)}\|^2_{\mathcal{H}_{K^{(2)}}} + \frac{\gamma_C}{(l+u)} \sum_{i=1}^{l+u} \left[f^{(1)}(x_i^{(1)}) - f^{(2)}(x_i^{(2)}) \right]^2$$

Here, μ is a real-valued parameter to balance data fitting in the two views, γ_1, γ_2 are regularization parameters for the RKHS norms in the two views, and γ_C is the coupling parameter that regularizes the pair towards compatibility using unlabeled data. It is easy to see that a representer theorem holds that expresses the minimizing pair $\left(f^{(1)*}(x^{(1)}), f^{(2)*}(x^{(2)})\right)$ in the following form:

$$\left(\sum_{i=1}^{l+u} \alpha_i K^{(1)}(x^{(1)}, x_i^{(1)}) \quad , \quad \sum_{i=1}^{l+u} \beta_i K^{(2)}(x^{(2)}, x_i^{(2)}) \right)$$

The $(l+u)$ dimensional expansion coefficient vectors α, β may be computed by solving the following coupled linear system:

$$\left[\frac{1}{l} J K_1 + \gamma_1 I + \frac{\gamma_C}{l+u} K_1 \right] \alpha - \frac{\gamma_C}{l+u} K_2 \beta = \frac{1}{l} Y$$

$$\left[\frac{\mu}{l} J K_2 + \gamma_2 I + \frac{\gamma_C}{l+u} K_2 \right] \beta - \frac{\gamma_C}{l+u} K_1 \alpha = \frac{\mu}{l} Y$$

where Y is a label vector given by $Y_i = y_i$ for $1 \le i \le l$ and $Y_i = 0$ for $l+1 \le i \le l+u$; J is a diagonal matrix given by $J_{ii} = |Y_i|$, and K_1, K_2 are gram matrices of the kernel functions $K^{(1)}, K^{(2)}$ over labeled and unlabeled examples.

When $\gamma_C = 0$, the system ignores unlabeled data and yields an uncoupled pair of solutions corresponding to supervised RLS. We also note a curious relationship over coefficients corresponding to unlabeled examples: $\gamma_1 \alpha_i = -\gamma_2 \beta_i$ for $l+1 \le i \le l+u$. The algorithm appears to work well in practice when orthogonality to the constant function is enforced over the data to avoid all unlabeled examples from being identically classified.

Working with the hinge loss, one can also extend SVMs in a similar manner. This has not been attempted in this paper.

Co-Laplacian RLS and Co-Laplacian SVM

The intuitions from the discussion concerning Figure 1(b) is to learn the pair $f = (f^{(1)}, f^{(2)})$ so that each function correctly classifies the labeled examples and is smooth with respect to similarity structures in both views. These structures may be encoded as graphs on which regularization operators may be defined and then combined to form a multi-view regularizer. The function pair is indirectly coupled through this regularizer.

We assume that for each view (indexed by $s = 1, 2$), we can construct a similarity graph whose adjacency matrix is $W^{(s)}$, where $W_{ij}^{(s)}$ measures similarity between $x_i^{(s)}$ and $x_j^{(s)}$. The Laplacian matrix of this graph is defined as $L^{(s)} = D^{(s)} - W^{(s)}$ where $D^{(s)}$ is the diagonal degree matrix $D_{ii}^{(s)} = \sum_j W_{ij}^{(s)}$. The graph Laplacian is a positive semi-definite operator on functions defined over vertices of the graph. It provides

the following smoothness functional on the graph:

$$\mathbf{g}^T L^{(s)} \mathbf{g} = \sum_{ij} (g_i - g_j)^2 W_{ij}^{(s)}$$

where $\mathbf{g}$ is a vector identifying a function on the graph whose value is g_i on node i. Other regularization operators can also be defined using the graph Laplacian (Kondor & Lafferty, 2003; Smola & Kondor, 2003; Belkin, Matveeva, & Niyogi, 2004).

One way to construct a multi-view regularizer is to simply take a convex combination $L = (1-\alpha)L^{(1)} + \alpha L^{(2)}$ where $\alpha \geq 0$ is a non-negative parameter which controls the influence of the two views. To learn the pair $f = (f^{(1)*}, f^{(2)*})$, we solve the following optimization problems for $s = 1, 2$ using squared loss or hinge loss:

$$f^{(s)*} = \underset{f^{(s)} \in \mathcal{H}_{K^{(s)}}}{\operatorname{argmin}} \frac{1}{l} \sum_{i=1}^{l} V(x_i^{(s)}, y_i, f^{(s)}) + \gamma_A^{(s)} \|f^{(s)}\|^2_{K^{(s)}} + \gamma_I^{(s)} \mathbf{f}^{(s)T} L \mathbf{f}^{(s)}$$

where $\mathbf{f}^{(s)}$ denotes the vector $\left(f^{(s)}(x_1^{(s)}), \ldots, f^{(s)}(x_{l+u}^{(s)})\right)^T$; and the regularization parameters $\gamma_A^{(s)}, \gamma_I^{(s)}$ control the influence of unlabeled examples relative to the RKHS norm.

The solutions to these optimization problems produce the recently proposed Laplacian SVM (for hinge loss) or Laplacian RLS (for squared loss) classifiers trained with the multi-view graph regularizer (Belkin, Niyogi & Sindhwani, 2004; Sindhwani, Niyogi & Belkin, 2005; Sindhwani, 2004). The resulting algorithms are termed Co-Laplacian SVM and Co-Laplacian RLS respectively.

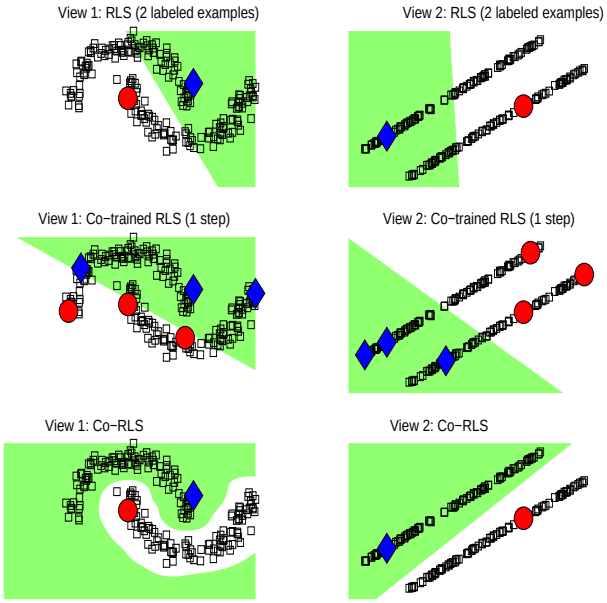
The solutions are obtained by training a standard SVM or RLS using the following modified kernel function:

$$\tilde{K}^{(s)}(x^{(s)}, z^{(s)}) = K^{(s)}(x^{(s)}, z^{(s)}) - \mathbf{k}^{\mathbf{T}}_{\mathbf{x}^{(s)}} (I + MG^{(s)})^{-1} M \mathbf{k}_{\mathbf{z}^{(s)}}$$

where $G^{(s)}$ is the gram matrix of the kernel function $K^{(s)}$; $\mathbf{k}_{\mathbf{x}^{(s)}}$ denotes the vector $\left(K^{(s)}(x_1^{(s)}, x^{(s)}), \ldots, K^{(s)}(x_n^{(s)}, x^{(s)})\right)^T$ and $M = \frac{\gamma_I^{(s)}}{\gamma_A^{(s)}} L$. See (Sindhwani, Niyogi & Belkin, 2005) for a derivation of this kernel.

When $\alpha = 0$ for view 1 or $\alpha = 1$ for view 2, the multi-view aspect is ignored and the pair consists of Laplacian SVM or Laplacian RLS in each view. When $\gamma_I = 0$, the unlabeled data is ignored and the pair consists of standard SVM or RLS classifiers.

Figure 2. Two-Moons-Two-Lines : RLS, Co-trained RLS and Co-RLS



The idea of combining graph regularizers and its connection to co-training has been briefly discussed in (Joachims, 2003) in the context of applying spectral graph transduction (SGT) in multi-view settings. However, unlike co-training, SGT does not produce classifiers defined everywhere in $X^{(1)}, X^{(2)}$ so that predictions cannot be made on novel test points. By optimizing in reproducing kernel Hilbert spaces defined everywhere, Co-Laplacian SVM and RLS can also extend beyond the unlabeled examples.

4. Experiments

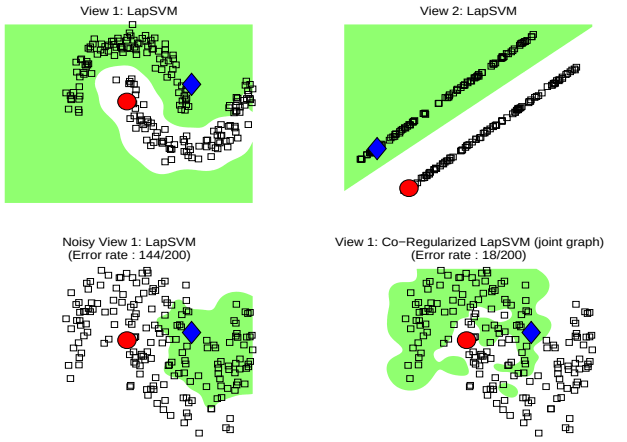
We performed experiments on a toy multi-view dataset and a hypertext document categorization task.

Two-Moons-Two-Lines Toy Example

Figure 2 and Figure 3 demonstrate Co-Regularization ideas on a toy dataset in which objects in two classes appear as two moons in one view and two oriented lines in another. Class conditional view independence is enforced by randomly associating points on one moon with points on one line, somewhat like the News 2×2 dataset in (Nigam & Ghani 2000). One example is labeled from each class and shown as the large colored diamond and circle; the other examples are unlabeled. We chose a Gaussian kernel for the two moons view and a linear kernel for the two lines view.

In the top panel of Figure 2, we see that a supervised Regularized least squares classifier is unable to

Figure 3. Two-Moons-Two-Lines : Laplacian SVM and Co-Laplacian SVM



produce reasonable classifiers with only 2 labeled examples. In the middle panel, we add two more labeled examples based on the most confident predictions (which are actually incorrect) of the supervised classifiers on the unlabeled data. The middle panel shows the classifiers obtained after 1 iteration of standard co-training with the boosted set of 4 labeled examples. Since greedy co-training does not revise conjectured labels, subsequent training fails to yield good classifiers in either view. By contrast, Co-Regularized Least squares classifiers, shown in panel 3, effectively use the unlabeled data in two views.

In the top panel of Figure 3, we show single-view semi-supervised learning with Laplacian SVMs in the two views. We then add noise to the two-moons view so that the two clusters are merged. This is shown in the bottom left panel. In this case, the unlabeled data fails to provide any structure for Laplacian SVM to exploit. However, when the joint graph laplacian is used, the rich structure in the two-lines view can be used to recover good decision boundaries in the two moons view. The bottom right panel shows the boundaries constructed by Co-Laplacian SVM.

Hypertext Categorization

We considered the WebKB hypertext categorization task studied in (Blum & Mitchell, 1998; Joachims, 2003; Nigam & Ghani 2000). There are 1051 web documents belonging to two classes: *course* or *non-course* from four universities. Only 12 examples are labeled. The two views are the textual content of a webpage (which we will call *page* representation) and the anchortext on links on other webpages pointing to the webpage (*link* representation).

The data was preprocessed into 3000 features for the page-view and 1840 features for the link view using the Rainbow software (McAllum, 1996). We used linear kernels for both views. We also considered a page+link representation with concatenated features.

The performance of several methods as measured by mean precision-recall breakeven point (PRBEP) is tabulated in Table 1. These methods are (a) RLS, SVM on fully labeled data sets and with 12 randomly chosen labeled examples; (b) single-view semi-supervised methods: SGT (Joachims, 2003), TSVM (Joachims, 1999), Laplacian SVM, Laplacian RLS (Belkin, Niyogi & Sindhwani, 2004; Sindhwani, Niyogi & Belkin, 2005); (c) multi-view semi-supervised methods: Co-RLS, Co-trained RLS, Co-trained SVM, Co-LapRLS and Co-LapSVM. In Table 1, Co-LapRLS1, Co-LapSVM1 use $\alpha = 0.5$ to combine graph Laplacians in page and link views; and Co-LapRLS2, Co-LapSVM2 use the mean graph Laplacian over page, link and page+link views, to bias classifiers in each view. The performance of supervised classifiers with full labels (RLS (full) and SVM (full)) is the mean PRBEP for 10-fold cross-validation. For all other methods, we average over random choices of 12 labeled examples (making sure that each class is sampled at least once) and measure the mean PRBEP evaluated over the remaining 1039 examples. We avoided the model selection issue due to the small size of the labeled set and chose best parameters over a small range of values.

The results in table 1 suggest that Co-LapSVM and Co-LapRLS are able to effectively use unlabeled examples in the two views. The link and page classifiers using 12 labeled examples, 1039 unlabeled examples and multi-view regularizers match the performance of supervised classifiers with access to all the labels. We also see that Co-RLS outperforms Co-trained RLS. In Table 2, we report the performance of Co-Laplacian SVM (using the mean graph Laplacian over the page, link and page+link views) in classifying unlabeled and test web-documents of four universities. The high correlation between performance on unlabeled and unseen test examples suggests that the method provides good extension outside the training set.

5. Conclusion

We have proposed extensions of regularization algorithms in a setting where unlabeled examples are easily available in multiple views. The algorithms provide natural extensions for SVM and RLS in such settings. We plan to further investigate the properties of these algorithms and benchmark them on real world tasks.

Table 1. Mean precision-recall breakeven points over unlabeled documents for a hypertext classification task.

View → Classifier ↓	link	page	page+ link
RLS (full)	94.4	94.0	97.8
SVM (full)	93.7	93.5	99.0
RLS (12)	72.0	71.6	78.3
SVM (12)	74.4	77.8	84.4
SGT	78.0	89.3	93.4
TSVM	85.5	91.4	92.2
LapRLS	80.8	89.0	93.1
LapSVM	81.9	89.5	93.6
Co-trained RLS	74.8	80.2	-
Co-RLS	80.8	90.1	-
Co-LapRLS1	93.1	90.8	90.4
Co-LapRLS2	94.4	92.0	93.6
Co-trained SVM	88.3	88.7	-
Co-LapSVM1	93.2	93.2	90.8
Co-LapSVM2	94.3	93.3	94.2

Table 2. Mean precision-recall breakeven points over test documents and over unlabeled documents (test , unlabeled)

University → View ↓	page+link	page	link
Cornell	91.6 , 90.9	88.9 , 88.8	88.2 , 88.7
Texas	94.8 , 95.5	91.6 , 92.4	90.9 , 93.5
Washington	94.7 , 94.9	94.0 , 93.9	93.7 , 92.4
Wisconsin	92.0 , 91.4	87.6 , 86.6	86.1 , 84.5

References

Abney, S. (2002) *Bootstrapping.* Proceedings of ACL 40

Blum A. & Mitchell T. (1998) *Combining Labeled and Unlabeled Data with Co-Training.* COLT

Belkin M., Niyogi P. & Sindhwani V. (2004) *Manifold Regularization : A Geometric Framework for Learning for Examples.* Technical Report, Dept. of Computer Science, Univ. of Chicago, TR-2004-06

Belkin M., Matveeva I. & Niyogi P. (2004) *Regression and Regularization on Large Graphs.* COLT

Brefeld, U. & Scheffer, T. (2004) *Co-EM Support Vector Learning.* ICML

Chapelle O. & Zien,A. (2005) *Semi-Supervised Classification by Low Density Separation.* Artificial Intelligence and Statistics, 2005

Collins, M. & Singer, Y. (1999) *Unsupervised Models for Named Entity Classification.* EMNLP/VLC-99

Dasgupta, S., Littman, M., & McAllester, D. (2001) *PAC Generalization Bounds for Co-Training* NIPS

Joachims T. (1999) *Transductive Inference for Text Classification using Support Vector Machines.* ICML

Joachims T. (2003) *Transductive Learning via Spectral Graph Partitioning.* ICML

Kondor I.R.& Lafferty, J. (2003) *Diffusion Kernels on Graphs and Other Discrete Input Spaces.* ICML

Nigam, K. & Ghani, R. (2001) *Kamal Nigam and Rayid Ghani. Analyzing the Effectiveness and Applicability of Co-training.* CIKM, pp. 86-93

McCallum, A.K. (1996) *Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering.* http://www.cs.cmu.edu/∼mccallum/bow

Poggio, T., & Girosi (1990) *Regularization Algorithms for Learning That Are Equivalent to Multilayer Networks.* Science 247:978-982

Schoelkopf, B. & Smola, A.J. (2002) *Learning with Kernels.* MIT Press, Cambridge, MA

Sindhwani, V. (2004) *Kernel Machines for Semi-supervised Learning.* Masters Thesis, University of Chicago

Sindhwani, V., Niyogi, P., & Belkin, M. (2005) *Beyond the point cloud: from Transductive to Semi-supervised Learning.* ICML

Smola A.J., & Kondor I.R (2003) *Kernels and Regularization on Graphs.* COLT

Vapnik V, (1998) *Statistical Learning Theory.* Wiley-Interscience

Yarowsky, D. (1995) *Unsupervised Word Sense Disambiguation Rivaling Supervised Methods.* ACL

Zhou, D., Bousquet, O., Lal, T. N., Weston J., & Schoelkopf, B. (2004) *Learning with Local and Global Consistency.* NIPS 16