Semi-Supervised Learning for Structured Output Variables

Author: Ulf Brefeld, Tobias Scheffer
Presentation: Yunsong Guo

Nov 7, 2006

The Problem

- Structured Learning with unlabeled data.
- How to utilize unlabeled data to improve performance?

Dasgupta et al. 2001

Theorem 1 *With probability at least $1-\delta$ over the choice of the sample S, we have that for all h_1 and h_2, if $\gamma_i(h_1,h_2,\delta) > 0$ for $1 \le i \le k$ then (a) f is a permutation and (b) for all $1 \le i \le k$,*

$$P(h_1 \neq i \mid f(y) = i, h_1 \neq \perp) \;\leq\; \frac{\widehat{P}(h_1 \neq i \mid h_2 = i, h_1 \neq \perp) + \epsilon_i(h_1,h_2,\delta)}{\gamma_i(h_1,h_2,\delta)}.$$

- h_1 predicts y from x_1, and h_2 predicts y from x_2.
- This theorem states in essence, if the sample size is large, and h_1 and h_2 (called partial prediction rules) largely agree on unlabeled data, then the disagreement is a good measure of error rate.
- This requires the assumption that x_1 and x_2 are conditionally independent given y.

Important Idea

- Dasgupta et al. (2001) give PAC bounds on the error of co-training.
- In terms of the disagreement rate of hypotheses on unlabeled data in two independent views.
- A corollary of their results that holds under general assumptions is:

$$Pr(f^1 \neq f^2) \geq \max\{Pr(err(f^1)), Pr(err(f^2))\}.$$

The Natural Idea

To minimize the error for labeled examples and maximize the agreement for unlabeled examples (among different views).

Normal Stuff

- Linear model: $\widehat{y} = argmax_{\overline{y} \in Y}\ f(x, \overline{y})$
- $f(x, y) =< w, \phi(x, y) >$
- Search for a minimizer for the empirical risk: $R_{emp}(f) = \sum_{i=1}^{n} \Delta(y_i, argmax_{\overline{y}} f(x_i, \overline{y}))$

Introduction of Views

- In co-learning, $\phi(x, y)$ are decomposed into disjoint sets $\phi^0(x, y)$ and $\phi^1(x, y)$.
- The spaces spanned are called views.
- For example, in hypertext classification we have two natural views on a page, either by the contained text or by the anchor text of its inbound links.
- The representation in each view has to be sufficient for the decoding.

3 Problems

1. Multi-Class Classification
2. Label Sequence Learning
3. Natural Language Parsing

Co-Support Vector Learning

- Large margin approach.
- Formulated 6 optimization problems incrementally.
- First 4 are more algorithmic, while the last 2 dual representations are for computational conveniences.

Co-Support Vector Learning

Optimization Problem 1 *Given n labeled examples; over all* $\mathbf{w}$ *minimize* $\frac{1}{2}\|\mathbf{w}\|^2$ *subject to the constraints* $\forall_{i=1}^n, \forall_{\bar{\mathbf{y}} \neq \mathbf{y}_i} \langle \mathbf{w}, \Phi(\mathbf{x}_i, \mathbf{y}_i) - \Phi(\mathbf{x}_i, \bar{\mathbf{y}}) \rangle \geq 1$.

Optimization Problem 2 *Given n labeled examples, let $C > 0$ and $r = 1, 2$; over all* $\mathbf{w}$ *and ξ_i minimize* $\frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{r}\sum_{i=1}^n \xi_i^r$ *subject to the constraints* $\forall_{i=1}^n \xi_i \geq 0$ *and* $\forall_{i=1}^n, \forall_{\bar{\mathbf{y}} \neq \mathbf{y}_i} \langle \mathbf{w}, \Phi(\mathbf{x}_i, \mathbf{y}_i) - \Phi(\mathbf{x}_i, \bar{\mathbf{y}}) \rangle \geq 1 - \xi_i$.

Co-Support Vector Learning

- Want to: integrate a loss function Δ into structured optimization problems.
- Two possible approaches: margin re-scaling (Taskar et al, 04) and slack re-scaling (Tsochantaridis et al, 05).
- Use slack re-scaling in this paper because with re-scaled slack variables, $\sum \xi_i$ still bounds the empirical lost.

Co-Support Vector Learning

Optimization Problem 3 *Given n labeled examples, loss function $\Delta : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_0^+$, tradeoff $C > 0$, and $r = 1, 2$; over all $\mathbf{w}$ and ξ_i minimize $\frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{r}\sum_{i=1}^{n} \xi_i^r$ subject to the constraints $\forall_{i=1}^{n} \xi_i \geq 0$ and $\forall_{i=1}^{n}, \forall_{\bar{\mathbf{y}} \neq \mathbf{y}_i} \langle \mathbf{w}, \Phi(\mathbf{x}_i, \mathbf{y}_i) - \Phi(\mathbf{x}_i, \bar{\mathbf{y}}) \rangle \geq 1 - \frac{\xi_i}{\sqrt[r]{\Delta(\mathbf{y}_i, \bar{\mathbf{y}})}}$.*

Co-Support Vector Learning: Incorporate Unlabeled Data

According to the consensus maximizing principle, need to minimize number of errors for labeled examples and disagreement for unlabeled examples.

- Use the prediction of the other view as the "right" label.

Co-Support Vector Learning: Incorporate Unlabeled Data

$$f^{v}(\mathbf{x}_i, \hat{\mathbf{y}}_i^{\bar{v}}) - \max_{\bar{\mathbf{y}} \neq \mathbf{y}_i} f^{v}(\mathbf{x}_i, \bar{\mathbf{y}}) = \gamma_i^{v} \geq 1$$

Optimization Problem 4 *Given n labeled examples and m unlabeled examples, loss function Δ, let $C, C_u > 0$, $r = 1, 2$, and $v = 0, 1$; over all $\mathbf{w}$ and ξ minimize* $\frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{r}\left(\sum_{i=1}^{n} \xi_i^r + C_u \sum_{i=n+1}^{n+m} (\min\{\gamma_i^{\bar{v}}, 1\})\xi_i^r\right)$ *subject to the constraints* $\forall_{i=1}^{n+m} \xi_i \geq 0$ *and* $\forall_{i=1}^{n}, \forall_{\bar{\mathbf{y}} \neq \mathbf{y}_i} \langle \mathbf{w}, \Phi(\mathbf{x}_i, \mathbf{y}_i) - \Phi(\mathbf{x}_i, \bar{\mathbf{y}}) \rangle \geq 1 - \frac{\xi_i}{\sqrt[r]{\Delta(\mathbf{y}_i, \bar{\mathbf{y}})}}$, $\forall_{i=n+1}^{n+m}, \forall_{\bar{\mathbf{y}} \neq \mathbf{y}^{\bar{v}}} \langle \mathbf{w}, \Phi(\mathbf{x}_i, \mathbf{y}_i^{\bar{v}}) - \Phi(\mathbf{x}_i, \bar{\mathbf{y}}) \rangle \geq 1 - \frac{\xi_i}{\sqrt[r]{\Delta(\mathbf{y}_i^{\bar{v}}, \bar{\mathbf{y}})}}$.

Co-Support Vector Learning: Dual Representation

- Introduce Lagrangian multipliers.
- Then take derivative of Lagrangian with respect to weight vector w.
- This leads to the dual representation.

Co-Support Vector Learning: Dual Representation

Optimization Problem 5 *Given n labeled and m unlabeled examples, loss function* Δ, $C, C_u > 0$; *over all* $\alpha_{i,\bar{\mathbf{y}}}$ *maximize*

$$\sum_{i=1}^{n+m} \sum_{\bar{\mathbf{y}} \neq \mathbf{y}_i} \alpha_{i,\bar{\mathbf{y}}} - \frac{1}{2} \sum_{i,j=1}^{n+m} \sum_{\substack{\bar{\mathbf{y}} \neq \mathbf{y}_i \\ \bar{\mathbf{y}}' \neq \mathbf{y}_j}} \alpha_{i,\bar{\mathbf{y}}} \alpha_{j,\bar{\mathbf{y}}'} K\left((\mathbf{x}_i, \bar{\mathbf{y}}), (\mathbf{x}_j, \bar{\mathbf{y}}')\right)$$

subject to the constraints $\forall_{i=1}^{n} \sum_{\bar{\mathbf{y}} \neq \mathbf{y}_i} \frac{\alpha_{i,\bar{\mathbf{y}}}}{\Delta(\mathbf{y}_i, \bar{\mathbf{y}})} \leq C$, $\forall_{i=n+1}^{n+m} \sum_{\bar{\mathbf{y}} \neq \mathbf{y}_i^{\bar{v}}} \frac{\alpha_{i,\bar{\mathbf{y}}}}{\Delta(\mathbf{y}_i^{\bar{v}}, \bar{\mathbf{y}})} \leq (\min\{\gamma_i^{\bar{v}}, 1\}) C_u C$, *and* $\forall_{i=1}^{n+m} \forall_{\bar{\mathbf{y}} \neq \mathbf{y}_i} \alpha_{i,\bar{\mathbf{y}}} \geq 0$.

Co-Support Vector Learning: Dual Representation

Optimization Problem 6 *Given n labeled and m unlabeled examples, loss function* Δ, $C, C_u > 0$; *over all* $\alpha_{i,\bar{\mathbf{y}}}$ *maximize*

$$\sum_{i=1}^{n+m} \sum_{\bar{\mathbf{y}} \neq \mathbf{y}_i} \alpha_{i,\bar{\mathbf{y}}} - \frac{1}{2} \sum_{i,j=1}^{n+m} \sum_{\substack{\bar{\mathbf{y}} \neq \mathbf{y}_i \\ \bar{\mathbf{y}}' \neq \mathbf{y}_j}} \alpha_{i,\bar{\mathbf{y}}} \alpha_{j,\bar{\mathbf{y}}'} K'((\mathbf{x}_i, \bar{\mathbf{y}}), (\mathbf{x}_j, \bar{\mathbf{y}}'))$$

subject to the constraints $\forall_{i=1}^{n+m} \forall_{\bar{\mathbf{y}} \neq \mathbf{y}_i} \alpha_{i,\bar{\mathbf{y}}} \geq 0.$

$$K'((\mathbf{x}_i, \bar{\mathbf{y}}), (\mathbf{x}_j, \bar{\mathbf{y}}')) = K((\mathbf{x}_i, \bar{\mathbf{y}}), (\mathbf{x}_j, \bar{\mathbf{y}}')) + \delta_{i\bar{\mathbf{y}}, j\bar{\mathbf{y}}'}$$

Co-Support Vector Learning: Algorithm

Algorithm 1 CoSVM Optimization Algorithm

Input: i-th unlabeled example $\mathbf{x}_i$, $S^0_{j\neq i}$, $S^1_{j\neq i}$, C, C_u, norm r, repetitions r_{max}.

Set $S_i^0 = S_i^1 = \emptyset$, $\alpha^0_{i,\mathbf{y}} = \alpha^1_{i,\mathbf{y}} = 0$ for all $\mathbf{y} \in \mathcal{Y}$
repeat
  for each view $v = 0, 1$ **do**
    $\hat{\mathbf{y}}^v = \text{argmax}_{\mathbf{y}} \langle \mathbf{w}^v, \Phi^v(\mathbf{x}_i, \mathbf{y}) \rangle$
    $\bar{\mathbf{y}}^v = \text{argmax}_{\mathbf{y}\neq\hat{\mathbf{y}}^v} (1 - \langle \mathbf{w}^v, \Phi^v_{i,\hat{\mathbf{y}}^v,\mathbf{y}} \rangle) \sqrt[r]{\Delta(\hat{\mathbf{y}}^v, \mathbf{y})}$
    $\xi_i^v = \max_{\mathbf{y}\in S_i^v} \{(1 - \langle \mathbf{w}, \Phi^v_{i,\hat{\mathbf{y}}^v,\mathbf{y}} \rangle) \sqrt[r]{\Delta(\hat{\mathbf{y}}^v, \mathbf{y})}\}$
    $\gamma^v = f^v(\mathbf{x}_i, \hat{\mathbf{y}}^v) - f^v(\mathbf{x}_i, \bar{\mathbf{y}}^v)$
  end for
  if $[[\hat{\mathbf{y}}^0 \neq \hat{\mathbf{y}}^1]] \vee [[\langle \mathbf{w}^v, \Phi^v_{i,\hat{\mathbf{y}}^v,\bar{\mathbf{y}}^v} \rangle < 1 - \frac{\xi_i^v}{\sqrt[r]{\Delta(\hat{\mathbf{y}}^v,\bar{\mathbf{y}}^v)}}]]$, $v = 0, 1$ **then**
    for each view $v = 0, 1$ **do**
      Substitute former target $\mathbf{y}_i^v = \hat{\mathbf{y}}^{\bar{v}}$
      if $[[\hat{\mathbf{y}}^0 \neq \hat{\mathbf{y}}^1]]$ **then**
        $S_i^v = S_i^v \cup \{\hat{\mathbf{y}}^v\}$
      else
        $S_i^v = S_i^v \cup \{\bar{\mathbf{y}}^v\}$
      end if
      Optimize $\alpha^v_{i,\bar{\mathbf{y}}}$ over S_i^v with $S^v_{j\neq i}$ fixed
      $\forall \bar{\mathbf{y}} \in S^v$ with $\alpha^v_{i,\bar{\mathbf{y}}} = 0$: $S_i^v = S_i^v \backslash \{\bar{\mathbf{y}}\}$
    end for
  end if
until consensus or r_{max} repetitions

Output: Optimized α_i^0 and α_i^1, sets S_i^0 and S_i^1

Experiment: TSVM

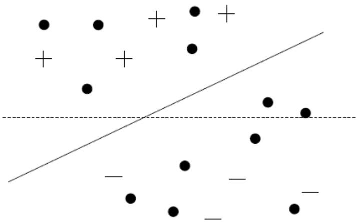

Figure 2: The maximum margin hyperplanes. Positive/negative examples are marked as $+/-$, test examples as dots. The dashed line is the solution of the inductive SVM. The solid line shows the transductive classification.

Experiment: TSVM

1. Use normal SVM on training set.
2. Predict on test set, get y^*.
3. Solve the following optimization problem:

OP 2 (Transductive SVM (non-sep. case))
Minimize over $(y_1^*, ..., y_n^*, \vec{w}, b, \xi_1, ..., \xi_n, \xi_1^*, ..., \xi_k^*)$:

$$\frac{1}{2}||\vec{w}||^2 + C\sum_{i=0}^{n}\xi_i + C^*\sum_{j=0}^{k}\xi_j^*$$

subject to:

$$\forall_{i=1}^{n} : y_i[\vec{w} \cdot \vec{x}_i + b] \geq 1 - \xi_i$$
$$\forall_{j=1}^{k} : y_j^*[\vec{w} \cdot \vec{x}_j^* + b] \geq 1 - \xi_j^*$$
$$\forall_{i=1}^{n} : \xi_i > 0$$
$$\forall_{j=1}^{k} : \xi_j^* > 0$$

Experiment

Table 1. Error rates for the Cora data set.

	L:200			L:400		
	U:0	U:400	U:800	U:0	U:800	U:2000
SVM	46.74 ± 0.26	-	-	38.39 ± 0.22	-	-
TSVM	46.13 ± 0.41	48.54 ± 0.28	50.84 ± 0.30	37.65 ± 0.25	39.31 ± 0.45	42.72 ± 0.60
coSVM	$\mathbf{41.94 \pm 0.30}$	42.51 ± 0.33	41.52 ± 0.26	$\mathbf{32.80 \pm 0.22}$	32.79 ± 0.21	32.72 ± 0.26

Table 2. Token error for the Biocreative (BC) and Spanish news wire (SN) data sets.

		L:5		L:10		L:20	
		U:0	U:25	U:0	U:50	U:0	U:100
BC	HMM	17.98 ± 0.69	-	14.32 ± 0.53	-	12.31 ± 0.23	-
	SVM	10.27 ± 0.16	-	9.70 ± 0.07	-	9.47 ± 0.05	-
	coSVM	$\mathbf{9.71 \pm 0.07}$	$\mathbf{9.54 \pm 0.08}$	$\mathbf{9.48 \pm 0.05}$	9.51 ± 0.05	9.4 ± 0.05	9.37 ± 0.06
SN	HMM	23.59 ± 2.00	-	20.04 ± 1.27	-	15.31 ± 0.78	-
	SVM	10.95 ± 0.18	-	9.98 ± 0.09	-	8.97 ± 0.08	-
	coSVM	13.86 ± 0.78	$\mathbf{10.28 \pm 0.14}$	11.26 ± 0.13	$\mathbf{9.60 \pm 0.11}$	11.73 ± 0.43	8.99 ± 0.09

Table 3. F1 scores for the wall street journal (WSJ) and the Negra (NEG) corpus.

		L:4		L:40		
		U:0	U:80	U:0	U:80	U:200
WSJ	SVM	45.40 ± 0.61	-	71.73 ± 0.29	-	-
	coSVM	$\mathbf{47.92 \pm 0.59}$	$\mathbf{48.23 \pm 0.55}$	$\mathbf{73.85 \pm 0.24}$	$\mathbf{74.07 \pm 0.25}$	$\mathbf{75.01 \pm 0.31}$
NEG	SVM	47.58 ± 0.37	-	63.70 ± 0.29	-	-
	coSVM	$\mathbf{48.81 \pm 0.37}$	$\mathbf{49.46 \pm 0.33}$	$\mathbf{64.94 \pm 0.27}$	$\mathbf{65.13 \pm 0.25}$	$\mathbf{65.70 \pm 0.25}$

Experiment

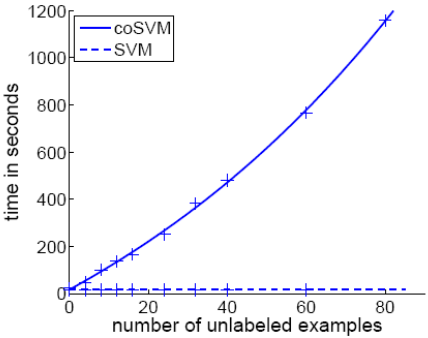

Figure 2. Execution time.

Conclusion

1. Devised a semi-supervised variant of SVM for structured learning.
2. Devised 1-norm and 2-norm optimization problems that allow to use arbitrary feature mappings.
3. Better performance of coSVM comes with the price of longer execution time.