Two-view Transductive Support Vector Machines

A Multi-view Semi-supervised Learning Algorithm

Guangxia Li Steven C. H. Hoi Kuiyu Chang

Nanyang Technological University

Motivation

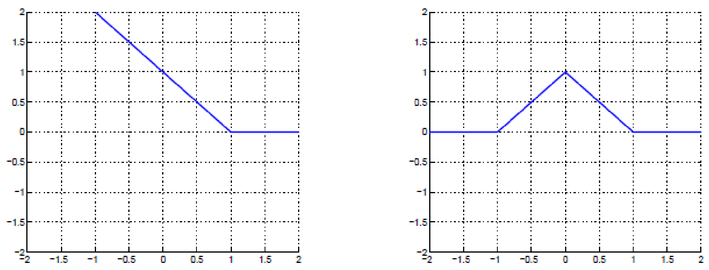
1. Scarce labeled data, abundant unlabeled data
   Solution:
   - Semi-supervised learning
2. For some problems, two or more sources of the inputs are given with the same output
   Solution:
   - Multi-view learning
3. A novel multi-view semi-supervised learning algorithm — two-view transductive SVM that takes advantage of both
   - the abundant amount of unlabeled data
   - and their multiple representations

   to improve the classification result

Semi-supervised Learning

- A set of l labeled examples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{l}$, and a set of u unlabeled examples $\{\boldsymbol{x}_i\}_{i=l+1}^{l+u}$, $\boldsymbol{x}_i \in \mathbb{R}^n$, $y_i \in \{-1, +1\}$
- The transductive SVM solves the following problem

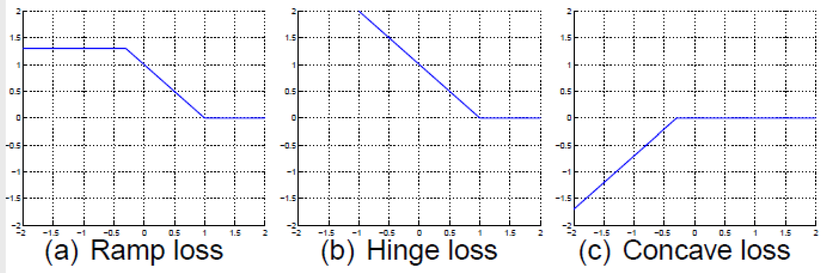
$$\min \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{l} L(y_i f_\theta(\boldsymbol{x}_i)) + C^* \sum_{i=l+1}^{l+u} L(|f_\theta(\boldsymbol{x}_i)|)$$



(a) Hinge loss (b) Symmetric hinge loss

- The objective function is hard to optimize since it is non-convex
- CCCP transductive SVM (CCCP TSVM), proposed by Collobert et al. [Collobert, 2006]
- A special loss function — "ramp loss", is applied to the unlabeled data
- The "ramp loss" $R_s(\cdot)$ can be expressed as the sum of a hinge loss $L(\cdot)$ and a concave loss $L_s(\cdot)$

$$R_s(\cdot) = \min(1-s, \max(0, 1-\cdot)) = L(\cdot) + L_s(\cdot)$$



(a) Ramp loss (b) Hinge loss (c) Concave loss

- Train a SVM using the hinge loss for labeled data, and the ramp loss for unlabeled data

Multi-view Learning

For some problems, there may exist multiple perspectives, so called views, of each data sample

Example

For text classification

- The typical view contains a large number of raw content features such as term frequency
- Another view may contain a small but highly-informative number of domain specific features

- Can we use the multiple representations of the same problem to improve the performance of classifiers?
- SVM-2K, proposed by Farquhar et al. [Farquhar, 2005]

[Farquhar, 2005] J. D. R. Farquhar, D. R. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak, Two view learning: SVM-2K, theory and practice, *NIPS*, 2005.

- Two SVM classifiers are created from each view of the same problem
- Each classifier is optimized to maximize the overall consensus of their predictions

$$|f_\theta^A(\boldsymbol{x}_i^A) - f_\theta^B(\boldsymbol{x}_i^B)| \le \eta_i + \varepsilon$$

Our Approach: Two-view Transductive SVM

- A multi-view semi-supervised learning problem contains
  - A set of l labeled examples $\{(\boldsymbol{x}_i^A, \boldsymbol{x}_i^B), y_i\}_{i=1}^{l}$, $y_i = \pm 1$
  - A set of u unlabeled examples $\{\boldsymbol{x}_i^A, \boldsymbol{x}_i^B\}_{i=l+1}^{l+u}$, $\boldsymbol{x}_i^{A/B} \in \mathbb{R}^n$
- For each view, we aim to find a decision function $f_\theta(\boldsymbol{x})$ as

$$f_\theta(\boldsymbol{x}) = \boldsymbol{w} \cdot \Phi(\boldsymbol{x}) + b$$

where $\theta = (\boldsymbol{w}, b)$ are the parameters of the model, and $\Phi(\cdot)$ is the feature map

- The optimization problem of two-view transductive SVM is

$$\begin{aligned}
\min_{\theta^{A/B}, \xi^{A/B}, \eta} \quad & \frac{1}{2}\|\boldsymbol{w}^A\|^2 + C^A \sum_{i=1}^{l} \xi_i^A + C^{*A} \sum_{i=l+1}^{l+2u} \xi_i^A + \sum_{i=l+1}^{l+2u} \rho_i^A y_i f_\theta^A(\boldsymbol{x}_i^A) + \\
& \frac{1}{2}\|\boldsymbol{w}^B\|^2 + C^B \sum_{i=1}^{l} \xi_i^B + C^{*B} \sum_{i=l+1}^{l+2u} \xi_i^B + \sum_{i=l+1}^{l+2u} \rho_i^B y_i f_\theta^B(\boldsymbol{x}_i^B) + D \sum_{i=1}^{l+2u} \eta_i \\
\text{s.t.} \quad & y_i f_\theta^{A/B}(\boldsymbol{x}_i^{A/B}) \ge 1 - \xi_i^{A/B} \\
& \xi_i^{A/B} \ge 0 \\
& \left| f_\theta^A(\boldsymbol{x}_i^A) - f_\theta^B(\boldsymbol{x}_i^B) \right| \le \eta_i + \varepsilon \\
& \eta_i \ge 0 \\
& \frac{1}{u} \sum_{i=l+1}^{l+u} f_\theta^{A/B}(\boldsymbol{x}_i^{A/B}) = \frac{1}{l} \sum_{i=1}^{l} y_i
\end{aligned}$$

- Applying the usual Lagrange multiplier technique, we have

$$\begin{aligned}
\min_{\tilde{\alpha}^{A/B}, \beta} \quad & \frac{1}{2} \sum_{i,j=0}^{l+2u} (y_i \tilde{\alpha}_i^A + \beta_i)(y_j \tilde{\alpha}_i^A + \beta_j) K_{ij}^A - \sum_{i=0}^{l+2u} g_i \tilde{\alpha}_i^A + \\
& \frac{1}{2} \sum_{i,j=0}^{l+2u} (y_i \tilde{\alpha}_i^B - \beta_i)(y_j \tilde{\alpha}_i^B - \beta_j) K_{ij}^B - \sum_{i=0}^{l+2u} g_i \tilde{\alpha}_i^B - \varepsilon \sum_{i=0}^{l+2u} \beta_i^2 \\
\text{s.t.} \quad & 0 \le \tilde{\alpha}_i^{A/B} \le C^{A/B} \qquad \forall 1 \le i \le l \\
& -\rho_i^{A/B} \le \tilde{\alpha}_i^{A/B} \le C^{*A/B} - \rho_i^{A/B} \qquad \forall l+1 \le i \le l+2u \\
& -D \le \beta_i \le D \qquad \forall 1 \le i \le l+2u \\
& \sum_{i=0}^{l+2u} (y_i \tilde{\alpha}_i^A + \beta_i) = 0 \\
& \sum_{i=0}^{l+2u} (y_i \tilde{\alpha}_i^B - \beta_i) = 0
\end{aligned}$$

- Applying the augmented Lagrangian technique to the minimization problem, we have

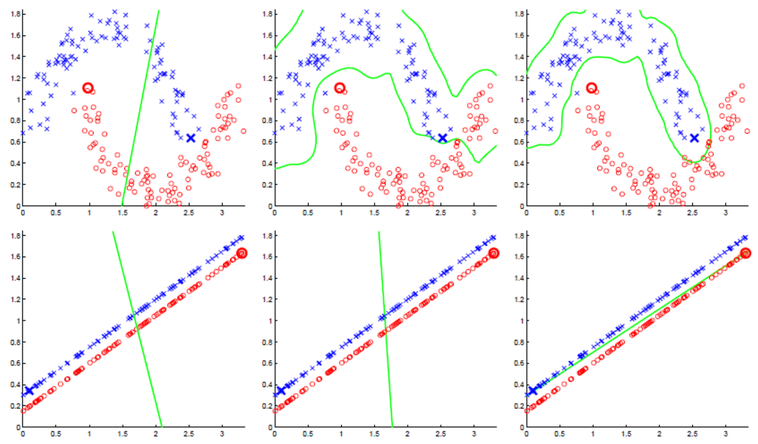
$$\begin{aligned}
\min_{\tilde{\alpha}^{A/B}, \beta} \quad & \frac{1}{2} \sum_{i,j=0}^{l+2u} (y_i \tilde{\alpha}_i^A + \beta_i)(y_j \tilde{\alpha}_i^A + \beta_j) K_{ij}^A - \sum_{i=0}^{l+2u} g_i \tilde{\alpha}_i^A - \varepsilon \sum_{i=0}^{l+2u} \beta_i^2 + \\
& \frac{1}{2} \sum_{i,j=0}^{l+2u} (y_i \tilde{\alpha}_i^B - \beta_i)(y_j \tilde{\alpha}_i^B - \beta_j) K_{ij}^B - \sum_{i=0}^{l+2u} g_i \tilde{\alpha}_i^B - \sum_{i=1}^{2} \lambda_i h_i + \frac{\mu}{2} \sum_{i=1}^{2} \|h_i\|^2 \\
\text{s.t.} \quad & 0 \le \tilde{\alpha}_i^{A/B} \le C^{A/B} \quad \forall 1 \le i \le l \\
& -\rho_i^{A/B} \le \tilde{\alpha}_i^{A/B} \le C^{*A/B} - \rho_i^{A/B} \quad \forall l+1 \le i \le l+2u \\
& -D \le \beta_i \le D \quad \forall 1 \le i \le l+2u
\end{aligned}$$

- Two-view Transductive SVM Outline

repeat
  Solve the following CCCP problem.
  repeat
    Solve minimization problem with fixed λ^k and μ^k.
    Update the lower and upper bounds of $\alpha_i^{A/B}$
  until The CCCP process converges
  Update the Lagrange multiplier λ by $\lambda_i^{k+1} = \lambda_i^k + \mu^k h_i^k$
  Update the panalty parameter μ by $\mu^{k+1} = \eta\mu^k$
until $\|\boldsymbol{h}^k\| \le \epsilon$

Experimental Results

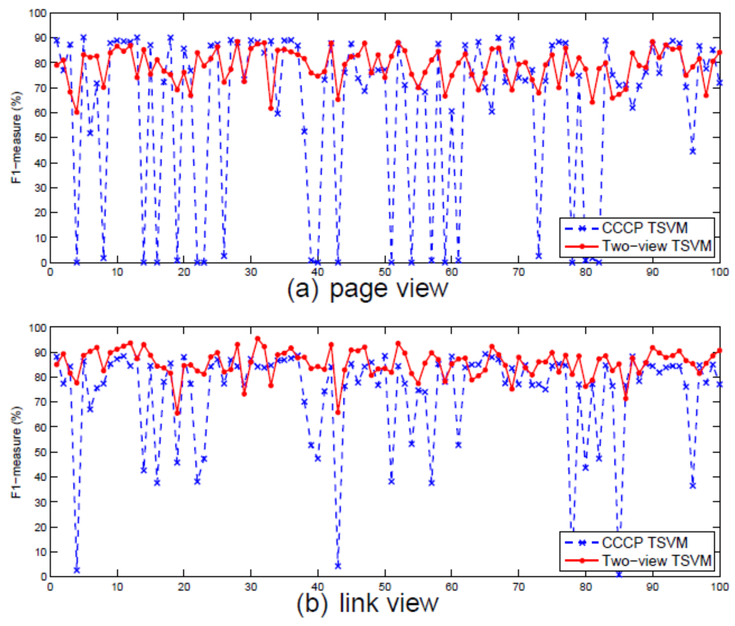
Toy Dataset



(a) Supervised SVM (b) CCCP TSVM (c) Two-view TSVM

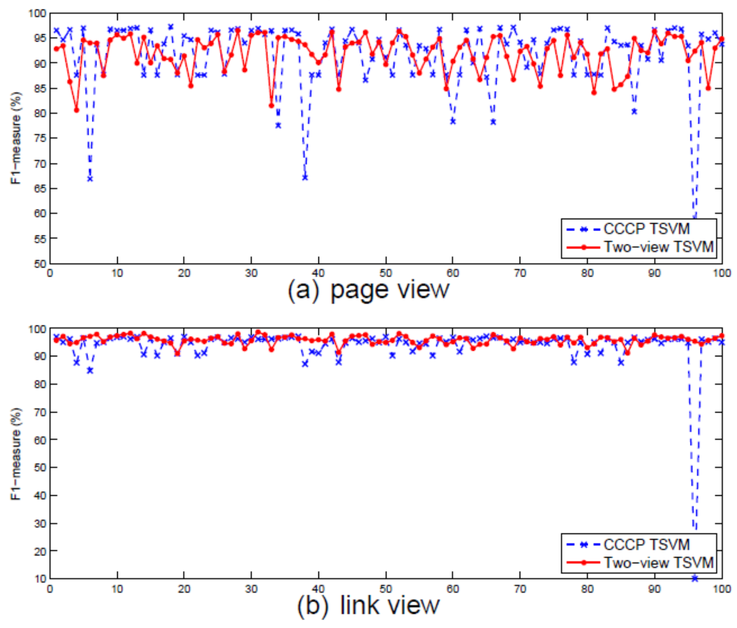
WebKB Course Dataset

•Variation of Positive Class's F1-measure



(a) page view (b) link view

•Variation of Negative Class's F1-measure



(a) page view (b) link view

Product Review Dataset

•Classification results showing mean accuracy (in percentage) and its standard deviation (in brackets)

Algorithm	Lexical view	Formal view	Hybrid view
SVM (20 labels)	57.63 (18.98)	69.00 (12.10)	-
SVM-2K (20 labels)	74.22 (5.24)	70.80 (4.76)	76.96 (4.77)
CCCP TSVM	76.30 (2.27)	74.39 (6.03)	-
Two-view TSVM	**79.17 (4.91)**	74.60 (5.20)	**80.43 (5.31)**

•Accuracy versus the number of labeled examples for CCCP TSVM and Two-view TSVM on the product review dataset