Resources of Transfer Learning

- Course
  - UC Berkeley http://www.cs.berkeley.edu/~russell/classes/cs294/f05/
  - UT Austin http://www.cs.utexas.edu/~lilyanam/TL/
- Workshop
  - Inductive Transfer : 10 Years Later NIPS 2005 Workshop http://iitrl.acadiau.ca/itws05/index.htm

A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data

Rie Kubota Ando
Tong Zhang
IBM T.J. Watson Research Center
Presented by Jiazhong Nie Feb.5 2009

Supervised learning

- Learn a predictor f:X->Y
- With some loss function L, the measure of a predicator is

$$R(f) = \mathbf{E}_{\mathbf{X},Y} L(f(\mathbf{X}), Y)$$

- Don't know the distribution (x,y)
  - empirical risk : use the sum of loss on training set $\{(\mathbf{X}_i, Y_i)\}$ generated independently instead of the expectation.

Supervised learning

- Learning: empirical risk minimization(ERM) in some hypothesis space H.

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} L(f(\mathbf{X}_i), Y_i).$$

- Hypothesis space is essential to the learning
- For multiple problems on the same the domain
  - Share information (parameter θ) among their hypothesis spaces.

Hypothesis spaces sharing

- For m learning problems indexed by $\ell \in \{1, \ldots, m\}$ the ERM is performed on each hypothesis space $\mathcal{H}_{\ell,\theta}$

$$\hat{f}_{\ell,\theta} = \arg \min_{f \in \mathcal{H}_{\ell,\theta}} \sum_{i=1}^{n_\ell} L(f(\mathbf{X}_i^\ell), Y_i^\ell)$$

$$\hat{\theta} = \arg \min_{\theta \in \Gamma} \left[r(\theta) + \sum_{l=1}^{m} O_l(X^l, Y^l, \theta) \right]$$

$$O_l(X, Y, \theta) = \min_{f \in H_{l,\theta}} \left(\sum_{i=1}^{n} L(f(X_i), Y_i) \right)$$

Example with the linear predictor

- Linear predictor with known feature map Φ

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{\Phi}(\mathbf{x})$$

- Introduce information shared by another feature mapping:

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}) + \mathbf{v}^T \mathbf{\Psi}_\theta(\mathbf{x})$$

- A simple linear form of $\mathbf{\Psi}_\theta(\mathbf{x}) = \Theta \mathbf{\Psi}(\mathbf{x})$,where $\mathbf{\Psi}$ is known

$$f_\Theta(\mathbf{w}, \mathbf{v}; \mathbf{x}) = \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}) + \mathbf{v}^T \Theta \mathbf{\Psi}(\mathbf{x})$$

linear predictor

- Hypothesis space

$$\mathcal{H}_\Theta = \left\{ \mathbf{w}^T \mathbf{\Phi}(\mathbf{x}) + \mathbf{v}^T \Theta \mathbf{\Psi}(\mathbf{x}) : \|\mathbf{w}\|_2 \le \frac{A}{\sup_{\mathbf{x}} \|\mathbf{\Phi}(\mathbf{x})\|_2}, \|\mathbf{v}\|_2 \le \frac{B}{\sup_{\mathbf{x}} \|\mathbf{\Psi}(\mathbf{x})\|_2} \right\}$$

$$\Gamma = \{\Theta \in R^{h \times p} : \Theta\Theta^T = I_{h \times h}\},$$

- Estimation
  - g(w,v) is some regularization condition of w,v.

$$[\{\hat{\mathbf{w}}_\ell, \hat{\mathbf{v}}_\ell\}, \hat{\Theta}] = \arg\min_{\{\mathbf{w}_\ell, \mathbf{v}_\ell\}, \Theta} \left[r(\Theta) + \sum_{\ell=1}^{m} \left(g(\mathbf{w}_\ell, \mathbf{v}_\ell) + \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} L(f_\Theta(\mathbf{w}_\ell, \mathbf{v}_\ell; \mathbf{X}_i^\ell), Y_i^\ell) \right) \right]$$

Optimization

- Objective function:

$$[\{\hat{\mathbf{w}}_\ell, \hat{\mathbf{v}}_\ell\}, \hat{\Theta}] = \underset{\{\mathbf{w}_\ell, \mathbf{v}_\ell\}, \Theta}{\arg\min} \sum_{\ell=1}^{m} \left(\frac{1}{n_\ell} \sum_{i=1}^{n_\ell} L((\mathbf{w}_\ell + \Theta^T \mathbf{v}_\ell)^T \mathbf{X}_i^\ell, Y_i^\ell) + \lambda_\ell \|\mathbf{w}_\ell\|_2^2 \right)$$
$$\text{s.t.} \quad \Theta\Theta^T = I_{h\times h} \ ,$$

- Introduce $\mathbf{u}_\ell = \mathbf{w}_\ell + \Theta^T \mathbf{v}_\ell$.

$$[\{\hat{\mathbf{u}}_\ell, \hat{\mathbf{v}}_\ell\}, \hat{\Theta}] = \underset{\{\mathbf{u}_\ell, \mathbf{v}_\ell\}, \Theta}{\arg\min} \sum_{\ell=1}^{m} \left(\frac{1}{n_\ell} \sum_{i=1}^{n_\ell} L(\mathbf{u}_\ell^T \mathbf{X}_i^\ell, Y_i^\ell) + \lambda_\ell \|\mathbf{u}_\ell - \Theta^T \mathbf{v}_\ell\|_2^2 \right)$$
$$\text{s.t.} \quad \Theta\Theta^T = I_{h\times h} \ .$$

Optimization

- 1.Fix θ and v, optimize with respect to u
  - A convex optimization problem with a convex choice of L
- 2.Fix u, optimize with respect to v and θ,

$$[\{\hat{\mathbf{v}}_\ell\}, \hat{\Theta}] = \arg\min_{\{\mathbf{v}_\ell\},\Theta} \sum_\ell \lambda_\ell \|\hat{\mathbf{u}}_\ell - \Theta^T \mathbf{v}_\ell\|_2^2, \quad \text{s.t.} \quad \Theta\Theta^T = I_{h\times h}.$$

- With some algebra, we know:

$$\min_{\mathbf{v}_\ell} \|\hat{\mathbf{u}}_\ell - \Theta^T \mathbf{v}_\ell\|_2^2 = \|\hat{\mathbf{u}}_\ell\|_2^2 - \|\Theta\hat{\mathbf{u}}_\ell\|_2^2.$$

Optimization

- Respect to θ

$$\hat{\Theta} = \arg\max_{\Theta} \sum_{\ell=1}^{m} \lambda_\ell \|\Theta \hat{\mathbf{u}}_\ell\|_2^2, \quad \text{s.t.} \quad \Theta\Theta^T = I_{h\times h}.$$

Let $\mathbf{U} = [\sqrt{\lambda_1}\hat{\mathbf{u}}_1, \ldots, \sqrt{\lambda_m}\hat{\mathbf{u}}_m]$ be an $p \times m$ matrix, we have

$$\hat{\Theta} = \arg\max_{\Theta} \ \mathrm{tr}(\Theta \mathbf{U}\mathbf{U}^T \Theta^T), \quad \text{s.t.} \quad \Theta\Theta^T = I_{h\times h},$$

- The solution is
  - SVD decomposition of U $\mathbf{U} = V_1 D V_2^T$
  - $\hat{\Theta}$ is the first h rows of V_1^T

An extension

- Some dimensions of X are more related to each other
  - Group dimensions
  - Perform SVD locally on each group
- With feature $[\mathbf{X}^{\ell}_{i,t}]_{t=1,\ldots,G}$

$$[\{\hat{\mathbf{w}}_{\ell,t}, \hat{\mathbf{v}}_{\ell,t}\}, \{\hat{\Theta}_t\}] = \mathop{\arg\min}_{\{\mathbf{w}_{\ell,t}, \mathbf{v}_{\ell,t}\}, \{\Theta_t\}} \sum_{\ell=1}^{m} \left(\frac{1}{n_\ell} \sum_{i=1}^{n_\ell} L\Big(\sum_{t=1}^{G} (\mathbf{w}_{\ell,t} + \Theta_t^T \mathbf{v}_{\ell,t})^T \mathbf{X}^{\ell}_{i,t}, Y_i^{\ell}\Big) + \sum_{t=1}^{G} \lambda_{\ell,t} \|\mathbf{w}_{\ell,t}\|_2^2 \right),$$

$$\text{s.t. } \forall t \in \{1, \ldots, G\}: \quad \Theta_t \Theta_t^T = I_{h_t \times h_t}. \qquad (10)$$

Semi-supervised Learning

- Systematically create multiple prediction problems(auxiliary problem) from unlabeled data
- Learn a good structural parameter θ by ERM on auxiliary problems
- Learn a predicator by ERM using θ on the original problem

Auxiliary problem creation

- With the characteristic:
  - Automatic labeling
  - Relevancy
- Ex: Word tagging task
  - Predict the word strings
- Ex: Text Classification
  - Predict frequent words: divide the word in a document into two groups, predict the most frequent words in one group based on the other group

Auxiliary problem creation

- Predict the behavior of the target classifier

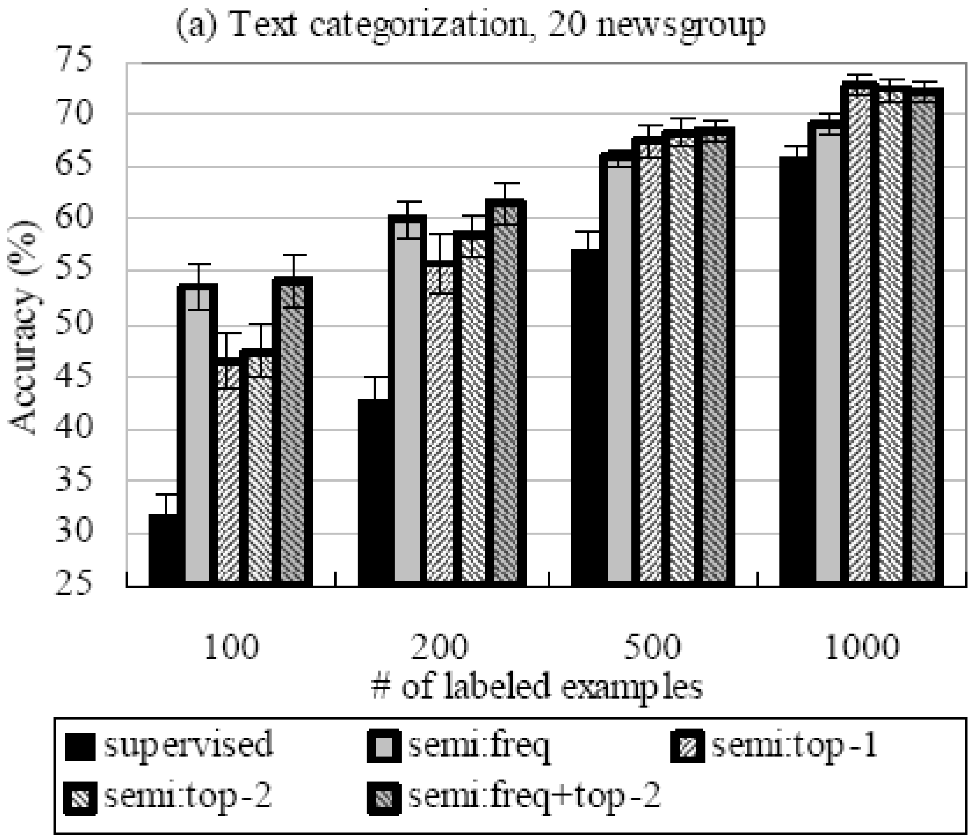
1. Train a classifier T_1 with labeled data Z for the target task, using feature map $\mathbf{\Phi}_1$.
2. Generate labeled data for auxiliary problems by applying T_1 to unlabeled data.
3. Learn structural parameter θ by performing joint ERM on the auxiliary problems, using only the feature map $\mathbf{\Phi}_2$.
4. Train a final classifier with labeled data Z, using θ computed above and some appropriate feature map $\mathbf{\Psi}$.

- Behavior:
  - Predict the prediction of classifier T
  - Predict the top-k choices of classifier T

Text Classification Experiment

- Corpus:
  - 20-newsgroup
- Feature:
  - Normalized word frequency vector
- Auxiliary problem:
  - Predicate the most frequent word based on one half of the words
  - Predicate top-K answers of the supervised classifier

Experiment result



- Significant improvements (22%) over the supervised method