Feature-aware Label Space Dimension Reduction for Multi-label Classification

Yao-Nan Chen (r99922008@csie.ntu.edu.tw) and *Hsuan-Tien Lin* (htlin@csie.ntu.edu.tw)

Department of Computer Science and Information Engineering, National Taiwan University

Multi-label Classification Setup

Which tags $\mathcal{Y}$ are associated with this picture $\mathbf{x}$?

$\mathcal{Y}$ = { building, taipei 101, ~~day view~~, night view, skyscraper, fireworks, ~~new york~~, fireworks, ~~car~~, ~~face~~, taipei world financial center, ~~university~~, etc.}

- Given: N examples $\left\{\left(\mathbf{x}_n \in \mathbb{R}^d, \mathcal{Y}_n \subseteq \{1, 2, \cdots, K\}\right)\right\}_{n=1}^{N}$
- Goal: classifier $g(\mathbf{x})$ that closely predicts the label-set $\mathcal{Y}$ associated with some unseen inputs $\mathbf{x}$, presumably by **exploiting hidden relations between labels**, e.g.
  - taipei 101 & taipei world financial center **highly correlated**
  - skyscraper **subset of** building
  - day view & night view **disjoint**

Label Space Dimension Reduction

$$\mathcal{Y} \subseteq \{1, 2, \cdots, K\} \text{ equivalent to } \mathbf{y} \in \{0, 1\}^K$$

- feature space dimension reduction: compress $\mathbf{x}$ to remove irrelevant, redundant (possibly related), or noisy information, **and achieve better efficiency & performance**
  - principal component analysis (PCA): linearly project $\mathbf{x}$ to $\mathbf{w}_m^T\mathbf{x}$ with minimum projection error
  - canonical correlation analysis (CCA): linearly project $\mathbf{x}$ to $\mathbf{w}_m^T\mathbf{x}$ in order to maximize correlation with some $\mathbf{v}_m^T\mathbf{y}$
- **label space dimension reduction**: analogously, but **compress y instead**

1. **compress**: transform $\{(\mathbf{x}_n, \mathbf{y}_n)\}$ to $\{(\mathbf{x}_n, \mathbf{t}_n)\}$ with $\mathbf{t}_n = \text{compress}(\mathbf{y}_n) \in \mathbb{R}^M$ and $M \ll K$
2. **learn**: train some $\mathbf{r}(\mathbf{x})$ from $\{(\mathbf{x}_n, \mathbf{t}_n)\}$
3. **decompress**: $g(\mathbf{x}) = \text{decompress}(\mathbf{r}(\mathbf{x}))$

- compressive sensing (Hsu et al., NIPS 2009): linearly project $\mathbf{y}$ to $\mathbf{t}[m] = \mathbf{v}_m^T\mathbf{y}$ with random $\mathbf{v}_m$'s (for incoherence)
- principal label space transformation (PLST; Tai and Lin, NC 2012): linearly project $\mathbf{y}$ to $\mathbf{t}[m] = \mathbf{v}_m^T\mathbf{y}$ with minimum projection error (**sibling of PCA**)

Feature-Aware Label Space Dimension Reduction

- feature space dimension reduction

unsupervised (not using $\mathbf{y}$)	supervised (using $\mathbf{y}$)
PCA, locally linear embedding, etc.	CCA, sliced inverse regression, etc.

—**supervised generally better for learning from** compress($\mathbf{x}$) **to y**

- label space dimension reduction

feature-unaware (not using $\mathbf{x}$)	feature-aware (using $\mathbf{x}$)
PLST, compressive sensing, etc.	???

—can we improve PLST by **feature-aware** label space dimension reduction?

Conditional Principal Label Space Transformation

- idea 1: exploit dual role of CCA to be feature-aware

$$\begin{aligned} & \text{project } \mathbf{x} \text{ to } \mathbf{w}_m^T\mathbf{x} \text{ in order to maximize correlation with some } \mathbf{v}_m^T\mathbf{y} \\ \equiv\ & \text{project } \mathbf{y} \text{ to } \mathbf{v}_m^T\mathbf{y} \text{ in order to maximize correlation with some } \mathbf{w}_m^T\mathbf{x} \\ \approx\ & \text{project } \mathbf{y} \text{ to } \mathbf{v}_m^T\mathbf{y} \text{ in order to minimize difference to some } \mathbf{w}_m^T\mathbf{x} \end{aligned}$$

$$\text{proposed OCCA} : \min_{\mathbf{W},\mathbf{V}} \|\mathbf{X}\mathbf{W}^T - \mathbf{Y}\mathbf{V}^T\|_F^2, \text{ s.t. } \mathbf{V}\mathbf{V}^T = \mathbf{I}$$

  - project to easiest-by-linear-regression directions

- idea 2: keep benefits of PLST for compression

$$\text{existing PLST} : \min_{\mathbf{V}} \|\mathbf{Y} - \mathbf{Y}\mathbf{V}^T\mathbf{V}\|_F^2, \text{ s.t. } \mathbf{V}\mathbf{V}^T = \mathbf{I}$$

  - project to most representative directions

- proposed algorithm: **conditional principal label space transformation** (CPLST)

$$\min_{\mathbf{W},\mathbf{V}} \underbrace{\|\mathbf{X}\mathbf{W}^T - \mathbf{Y}\mathbf{V}^T\|_F^2}_{\text{learning error}} + \underbrace{\|\mathbf{Y} - \mathbf{Y}\mathbf{V}^T\mathbf{V}\|_F^2}_{\text{compression error}}, \text{ s.t. } \mathbf{V}\mathbf{V}^T = \mathbf{I}$$

  - theoretical guarantee (Tai and Lin, NC 2012): when using linear regression as $\mathbf{r}$,

$$\textbf{hamming loss} \leq \text{learning error} + \text{compression error}$$

  - algorithmic simplicity: closed-form optimal $\mathbf{V}$ contains top eigenvectors of

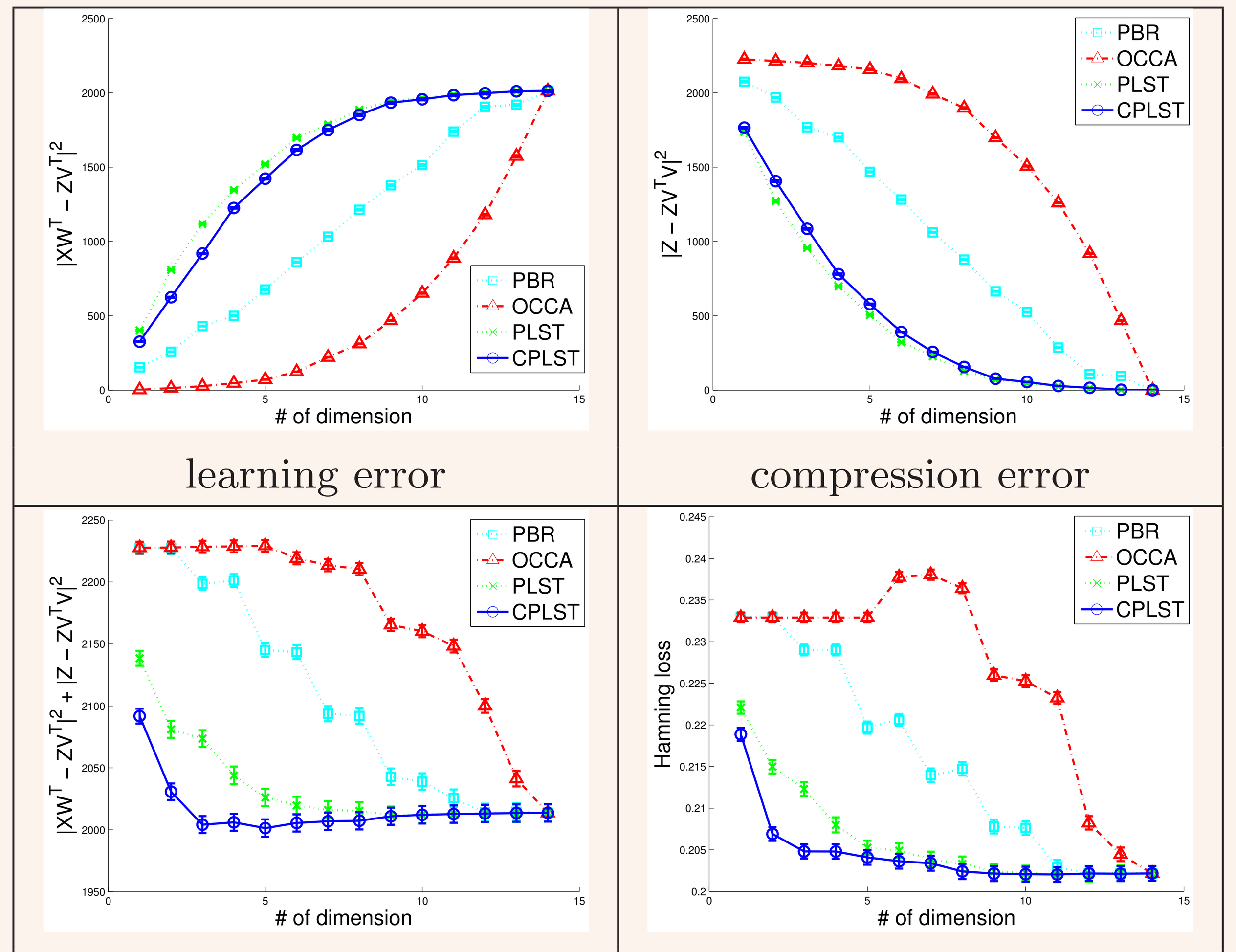
OCCA	PLST	CPLST
$\mathbf{Y}^T(\underbrace{\mathbf{X}\mathbf{X}^\dagger}_{\text{hat matrix}} - \mathbf{I})\mathbf{Y}$	$\mathbf{Y}^T\mathbf{Y}$	$\mathbf{Y}^T \underbrace{\mathbf{X}\mathbf{X}^\dagger}_{\text{hat matrix}} \mathbf{Y}$

(note: $\mathbf{Z}$, i.e. the mean-shifted $\mathbf{Y}$, is actually used for better projection)

  - physical meaning: exploit **conditional (feature-aware) correlations**
  - kernelization: replace linear regression with **kernel ridge regression** as $\mathbf{r}$

Experimental Results

on yeast data set:



learning error | compression error | learning+compression error | hamming loss

- PBR: baseline, with standard basis as $\mathbf{v}_m$
- OCCA: optimize learning error, but **worst in compression error**
- PLST: optimize compression error, but **worst in learning error**
- CPLST: optimize learning+compression error, and hence **best hamming loss**

on 8 benchmark data sets:

algorithms	CPLST vs. PLST + linear regression	CPLST vs. PLST + decision tree	kernel-CPLST vs. PLST + kernel ridge regression
$M = 20\%K$	**3 win**, 5 similar	**2 win**, 6 similar	**5 win**, 1 lose, 2 similar

CPLST **consistently better than or similar to PLST** across data & algorithms

Summary

Conditional Principal Label Space Transformation, which

- projects to **conditional** principal directions by combining ideas behind CCA (feature-aware) and PLST (optimal compression)
- can be **kernelized** for exploiting feature power
- achieves **better/similar** practical performance **consistently** when compared with the readily-strong PLST