

Extracting Shared Subspace for Multi-label Classification

Shuiwang Ji

Arizona State University

Joint work with Lei Tang, Shipeng Yu, and Jieping Ye

Multi-label Problems

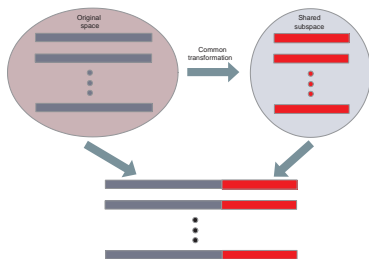
- Multi-label problems arise in many domains:
 - Text document categorization
 - Gene and protein function prediction
 - Image annotation
- One-against-rest scheme fails to exploit the correlations among labels
- Example:
 - Model the topics and authorship of documents
 - The topics and authors of documents are correlated, since a particular author may only write on certain topics
 - Model them jointly may reinforce each other

Problem Formulation

- Build one predictive function for each label
- A low-dimensional subspace is shared by all labels
- Each predictive function consists of two parts:
 - 1 First part depends on the original data representations
 - 2 Second part depends on the representations in the shared subspace
- Mathematically, the predictive function for the ℓ^{th} label is

$$f_{\ell}(x) = \mathbf{w}_{\ell}^T x + \mathbf{v}_{\ell}^T \Theta x, \quad (1)$$

where $\Theta \in \mathbb{R}^{r \times d}$ is common for all labels



Problem Formulation–Contd.

- The parameters $\{\mathbf{w}_\ell, \mathbf{v}_\ell\}_{\ell=1}^m$ and Θ are estimated by minimizing

$$\sum_{\ell=1}^m \left(\frac{1}{n} \sum_{i=1}^n L \left((\mathbf{w}_\ell + \Theta^T \mathbf{v}_\ell)^T x_i, y_i^\ell \right) + \alpha \|\mathbf{w}_\ell\|^2 + \beta \|\mathbf{w}_\ell + \Theta^T \mathbf{v}_\ell\|^2 \right), \quad (2)$$

subject to $\Theta \Theta^T = I$.

- When the least squares loss is used, the problem can be formulated as

$$\begin{aligned} \min_{U, V, \Theta} \quad & \frac{1}{n} \|XU - Y\|_F^2 + \alpha \|U - \Theta^T V\|_F^2 + \beta \|U\|_F^2 \quad (3) \\ \text{s. t.} \quad & \Theta \Theta^T = I, \end{aligned}$$

where $X = [x_1, \dots, x_n]^T$ is the data matrix, $U = [\mathbf{u}_1, \dots, \mathbf{u}_m]$, $V = [\mathbf{v}_1, \dots, \mathbf{v}_m]$, and $\mathbf{u}_\ell = \mathbf{w}_\ell + \Theta^T \mathbf{v}_\ell$.

The Main Theorem

Theorem

Let X , Y , and Θ be defined as above. Then the optimal Θ^* that solves the optimization problem in Eq. (3) can be obtained by solving the following trace maximization problem:

$$\begin{aligned} \max_{\Theta} \quad & \text{tr} \left(\left(\Theta S_1 \Theta^T \right)^{-1} \Theta S_2 \Theta^T \right) \\ \text{s. t.} \quad & \Theta \Theta^T = I, \end{aligned} \quad (4)$$

where M , S_1 , and S_2 are defined as:

$$M = \frac{1}{n} X^T X + (\alpha + \beta) I, \quad (5)$$

$$S_1 = I - \alpha M^{-1}, \quad (6)$$

$$S_2 = M^{-1} X^T Y Y^T X M^{-1}. \quad (7)$$

- Computationally expensive for high-dimensional data

An Efficient Algorithm for High-dimensional Data

- Compute the SVD of X as $X = U\Sigma V^T = U_1\Sigma_t V_1^T$
- Compute D_1 , D_2 , D , and \tilde{D} as

$$D_1 = \left(\frac{1}{n}\Sigma_t^2 + \beta I\right)^{-1}\Sigma_t \in \mathbb{R}^{t \times t}, \quad (8)$$

$$D_2 = \Sigma_t \left(\frac{1}{n}\Sigma_t^2 + (\alpha + \beta)I\right)^{-1} \in \mathbb{R}^{t \times t}, \quad (9)$$

$$D = (D_1 D_2^{-1})^{\frac{1}{2}} \in \mathbb{R}^{t \times t}, \quad (10)$$

$$\tilde{D} = D^{-1} D_1 = D_2 D. \quad (11)$$

- Compute the SVD of $C = Y^T U_1 \tilde{D}$ as $C = P_1 \Lambda P_2^T$
- Compute the QR decomposition of $V_1 D P_2$ as $V_1 D P_2 = QR$
- The rows of the optimal Θ^* are given by the first r columns of the matrix Q

Matrix	Size	Computation	Complexity
X	$n \times d$	SVD	$O(dn^2)$
C	$m \times t$	SVD	$O(tm^2)$
$V_1 D P_2$	$d \times t$	QR	$O(dt^2)$

- **Dimensionality Reduction**

- One set of variables are derived from the data and another set is derived from the class labels
- Canonical correlation analysis (CCA) and partial least squares (PLS) maximize the correlation and covariance of data and labels in the dimensionality-reduced space
- CCA reduces to linear discriminant analysis (LDA) for multi-class problems

- **Multi-task Learning** (Ando & Zhang, 2005)

- The input data for different tasks can be different
- The resulting optimization problem is non-convex even for convex loss functions
- Solved by an iterative procedure for a local solution

- **Multi-class Learning** (Amit *et al.*, 2007)

- Extracting shared structures in multi-class classification
- A low-rank transformation is computed to uncover the shared structures
- The problem is non-convex, it is relaxed to the convex trace norm constraint and solved by gradient-based optimization

The formulation is:

$$\begin{aligned} \min_{U, V, \Theta} \quad & \frac{1}{n} \|XU - Y\|_F^2 + \alpha \|U - \Theta^T V\|_F^2 + \beta \|U\|_F^2 \\ \text{s. t.} \quad & \Theta\Theta^T = I. \end{aligned} \quad (12)$$

- $\alpha = 0$: Equivalent to classical ridge regression
- $\beta = 0$: Reduces to the multi-task formulation (Ando & Zhang, 2005) where all tasks share the input data

The optimal Θ^* consists of the top eigenvectors of

$$S_1^{-1}S_2 = \left(\frac{1}{n}X^T X + \beta I \right)^{-1} X^T Y Y^T X \left(\frac{1}{n}X^T X + (\alpha + \beta)I \right)^{-1} \quad (13)$$

- $\alpha = +\infty$:
 - Equivalent to orthonormalized partial least squares (PLS)
 - When $Y Y^T$ is replaced by $Y(Y^T Y)^{-1}Y^T$, this problem reduces to canonical correlation analysis (CCA)
 - Equivalent to linear discriminant analysis (LDA) for multi-class problems, when

$$y_{ij} = \begin{cases} \sqrt{\frac{n}{n_j}} - \sqrt{\frac{n_j}{n}} & \text{if } y_i = j, \\ -\sqrt{\frac{n_j}{n}} & \text{otherwise,} \end{cases}$$

- $\beta = +\infty$: Closely related to the orthogonal centroid method (OCM)

- **ML_{LS}**: The proposed multi-label formulation based on least squares loss
- **CCA+Ridge**: CCA is applied first to reduce the data dimensionality before ridge regression is applied
- **CCA+SVM**: CCA is applied first to reduce the data dimensionality before linear SVM is applied
- **ASO_{SVM}**: The alternating structural optimization (ASO) algorithm (Ando & Zhang, 2005) with hinge loss
- **SVM**: Linear SVM is applied on each label using the one-against-rest scheme with the C value tuned using cross-validation
- **SVM_C**: Linear SVM is applied on each label using the one-against-rest scheme with C fixed to 1

All parameters are tuned using cross-validation.

Experiments—Performance

Table: Performance of the six methods on the Yahoo data sets (AUC, macro F1, and micro F1)

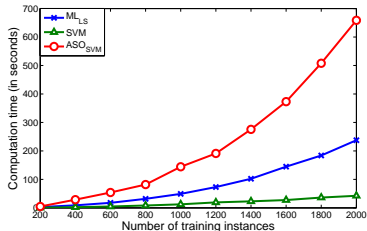
Algorithm	Arts	Business	Computer	Education	Entertainment	Health
ML_{LS}	0.7711	0.8348	0.7964	0.7753	0.8264	0.8483
CCA+Ridge	0.7571	0.8216	0.7932	0.7597	0.8043	0.8477
CCA+SVM	0.7519	0.8169	0.7774	0.7581	0.7965	0.8409
ASO_{SVM}	0.7678	0.8261	0.7847	0.7446	0.8207	0.8621
SVM_C	0.7674	0.8263	0.7943	0.7685	0.8155	0.8617
SVM	0.7668	0.8271	0.7940	0.7716	0.8177	0.8634
ML_{LS}	0.3583	0.3985	0.3219	0.3864	0.4874	0.5966
CCA+Ridge	0.3190	0.3779	0.2799	0.3602	0.4352	0.5431
CCA+SVM	0.3158	0.3758	0.3059	0.3618	0.4409	0.5338
ASO_{SVM}	0.3568	0.3736	0.2873	0.3262	0.4344	0.5814
SVM_C	0.3216	0.3533	0.2609	0.3588	0.4260	0.5632
SVM	0.3382	0.3677	0.2948	0.3830	0.4462	0.5705
ML_{LS}	0.4716	0.7645	0.5585	0.4899	0.5901	0.6809
CCA+Ridge	0.4444	0.7508	0.5414	0.4538	0.5506	0.6771
CCA+SVM	0.4524	0.7528	0.5394	0.4647	0.5498	0.6804
ASO_{SVM}	0.4449	0.7384	0.4305	0.4322	0.5605	0.6754
SVM_C	0.4449	0.7384	0.5275	0.4745	0.5413	0.6714
SVM	0.4574	0.7584	0.5458	0.4773	0.5701	0.6773

Experiments–Performance

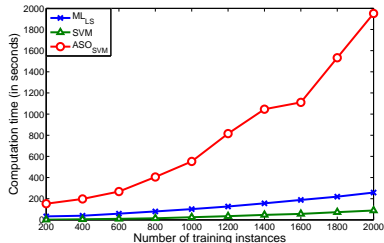
Table: Performance of the six methods on the Yahoo data sets (AUC, macro F1, and micro F1)

Algorithm	Recreation	Reference	Science	Social	Society
MLLS	0.8202	0.8358	0.8332	0.8320	0.7347
CCA+Ridge	0.8106	0.8260	0.8161	0.8189	0.7204
CCA+SVM	0.7896	0.8054	0.7946	0.7448	0.7007
ASO _{SVM}	0.8123	0.8340	0.8073	0.7942	0.7321
SVM _C	0.8092	0.8345	0.8277	0.8392	0.7308
SVM	0.8157	0.8327	0.8311	0.8377	0.7340
MLLS	0.4519	0.4208	0.4337	0.3680	0.3369
CCA+Ridge	0.4286	0.3330	0.3577	0.3279	0.2961
CCA+SVM	0.4331	0.3417	0.3650	0.3126	0.2996
ASO _{SVM}	0.4136	0.4116	0.3397	0.3017	0.3023
SVM _C	0.3852	0.3795	0.3770	0.3232	0.2919
SVM	0.4460	0.4005	0.4093	0.3380	0.3069
MLLS	0.5351	0.6020	0.5254	0.6606	0.4874
CCA+Ridge	0.5223	0.5336	0.4704	0.6607	0.4783
CCA+SVM	0.5159	0.5448	0.4815	0.6012	0.4690
ASO _{SVM}	0.4976	0.5580	0.4564	0.6492	0.4639
SVM _C	0.4797	0.5856	0.4774	0.6500	0.4569
SVM	0.5284	0.6002	0.5142	0.6573	0.4801

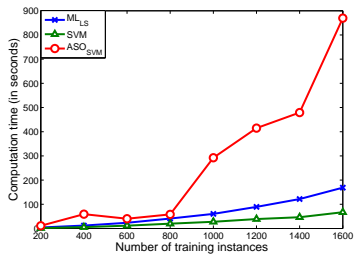
Experiments—Efficiency



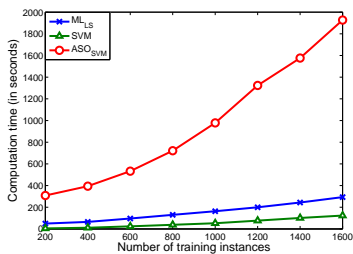
Health (with a fixed parameter)



Health (with parameter tuning)

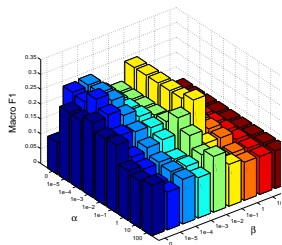


Science (with a fixed parameter)

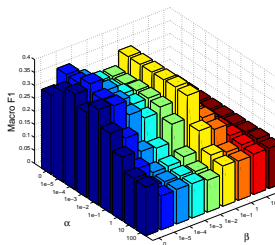


Science (with parameter tuning)

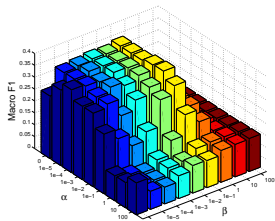
Figure: Comparison of computation time.



Arts



Recreation



Science

Figure: The change of macro F1 scores as the regularization parameters vary

THANKS!