Multi-Task Sparse Discriminant Analysis (MtSDA) with Overlapping Categories

Yahong Han and **Fei Wu**
College of Computer Science
Zhejiang University, China
{yahong,wufei}@zju.edu.cn

Jinzhu Jia
Department of Statistics
University of California,
Berkeley, CA 94720
jzjia@stat.berkeley.edu

Yueting Zhuang
College of Computer Science
Zhejiang University, China
yzhuang@zju.edu.cn

Bin Yu
Department of Statistics
and Department of EECS
University of California,
Berkeley, CA 94720
binyu@stat.berkeley.edu

Abstract

Multi-task learning aims at combining information across tasks to boost prediction performance, especially when the number of training samples is small and the number of predictors is very large. In this paper, we first extend the Sparse Discriminate Analysis (SDA) of Clemmensen *et al.*. We call this Multi-task Sparse Discriminate Analysis (MtSDA). MtSDA formulates multi-label prediction as a quadratic optimization problem whereas SDA obtains single labels via a nearest class mean rule. Second, we propose a class of equicorrelation matrices to use in MtSDA which includes the identity matrix. MtSDA with both matrices are compared with single-task learning (SVM and LDA+SVM) and multi-task learning (HSML). The comparisons are made on real data sets in terms of AUC and F-measure. The data results show that MtSDA outperforms other methods substantially almost all the time and in some cases MtSDA with the equicorrelation matrix substantially outperforms MtSDA with identity matrix .

Introduction

Multi-task learning attempts using the latent information hidden in related tasks. When applied appropriately, multi-task learning has advantages over traditional single task learning and therefore has many potential applications. Depending on how information is shared among the tasks, different algorithms have been devised. For example, hierarchical Bayesian modeling assumes that model parameters are shared by a common hyper prior (Yu, Tresp, and Schwaighofer 2005). For problems where the input lies in a high-dimensional space with a sparsity structure and only a few common important predictors are shared by tasks, regularized regression methods have been proposed to recover the shared sparsity structure across tasks (Lounici et al. 2009).

Multi-class (single-task) linear discriminant analysis (MLDA) is equivalent to the multi-response linear regression by optimal scoring (Hastie, Buja, and Tibshirani 1995). A benefit of performing MLDA via regression is the ability to introduce penalties to MLDA. In order to tackle problems of overfitting in situations of large numbers of highly correlated predictors, Hastie *et al.* (1995) introduced a quadratic penalty with a symmetric and positive definite matrix Ω into the objective function. Taking into account the ability of *elastic net* (Zou and Hastie 2005) which simultaneously conducts automatic variable selection and group selection of correlated variables, Clemmensen *et al.* (2008) formulated (single-task) MLDA as sparse discriminant analysis (SDA) by imposing both ℓ_1 and ℓ_2 norm regularization. However, it remained open how to extend penalized discriminant analysis to multi-task learning with overlapping categories.

In this paper, we are interested in the multi-task binary classification problem. Assume that we have a training set of n labeled data samples with J labels: $\{(x_i, y_i) \in R^p \times \{0,1\}^J, i = 1, 2, \ldots, n\}$, where $x_i = (x_{i1}, \ldots, x_{ip})' \in R^p$ represents the predictors for the ith data sample, and $y_i = (y_{i1}, \ldots, y_{iJ})' \in \{0,1\}^J$ is the corresponding response, $y_{ij} = 1$ if the ith data sample belongs to the jth category and $y_{ij} = 0$ otherwise. In multi-label setting, each data sample could belong to multiple categories and we want to learn a rule to predict the categories that a new unlabeled datum belongs to. Let $\mathbf{X} = (x_1, \ldots, x_n)'$ be the $n \times p$ training data matrix, and $\mathbf{Y} = (y_1, \ldots, y_n)'$ the corresponding $n \times J$ indicator response matrix. Unlike the traditional multi-task problem where each sample only belongs to a single category: $\sum_{j=1}^{J} y_{ij} = 1$, in overlapped multi-task learning we relax the constraint to $\sum_{j=1}^{J} y_{ij} \geq 0$.

We extend single-task SDA to the multi-task problem with a method we call multi-task sparse discriminant analysis (MtSDA). MtSDA uses a quadratic optimization approach for prediction of the multiple labels. In SDA the identity matrix is commonly used as the penalty matrix. Here we introduce a larger class of equicorrelation matrices with the identity matrix as a special case. We provide a theoretical result that indicates that an equicorrelation matrix has a grouping effect under some conditions.

Multi-task Sparse Discriminant Analysis (MtSDA)

In the non-overlapping category case, where each sample only belongs to one category, Clemmensen *et al.* (2008) proposed SDA based on formulating classification as regression as in Hastie *et al.* (1995). Define J mapping functions θ_j, $j \in \{1, 2, \ldots, J\}$, where

$$\theta_j : \{1, 2, \ldots, J\} \rightarrow R$$

assigns a score to each category. Denote by $\boldsymbol{\theta}$ the $J \times J$ score matrix, with (i, j) element $\boldsymbol{\theta}_{ij}$ defined as

$$\boldsymbol{\theta}_{ij} = \theta_j(i).$$

$\boldsymbol{\theta}_{ij}$ is the score assigned to the ith category by mapping θ_j. Denote by $\boldsymbol{\theta}_j$ the jth column of $\boldsymbol{\theta}$. Let $\boldsymbol{\Omega}$ be any $p \times p$ positive definite matrix and $\mathbf{I}_{J\times J}$ the $J \times J$ identity matrix. Denote by $\boldsymbol{\beta}_j \in R^p$ the coefficient vector from a linear regression of $\mathbf{Y}\boldsymbol{\theta}_j$ on $\mathbf{X}$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$ the $p \times J$ coefficient matrix. The SDA in (Clemmensen, Hastie, and Ersbøll 2008) is formulated as

$$\begin{aligned} \min_{\boldsymbol{\beta},\boldsymbol{\theta}} \frac{1}{n} \sum_{j=1}^{J} \left(||\mathbf{Y}\boldsymbol{\theta}_j - \mathbf{X}\boldsymbol{\beta}_j||_2^2 + \lambda_2 \boldsymbol{\beta}_j' \boldsymbol{\Omega} \boldsymbol{\beta}_j + \lambda_1 ||\boldsymbol{\beta}_j||_1 \right), \\ s.t. \qquad \frac{1}{n}(\mathbf{Y}\boldsymbol{\theta})'(\mathbf{Y}\boldsymbol{\theta}) = \mathbf{I}, \end{aligned} \tag{1}$$

where $||\mathbf{A}||_2^2 = \sum_{i,j} A_{i,j}^2$ is the ℓ_2 norm of matrix $\mathbf{A}$.

After all of the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ in (1) are obtained and a new unlabeled datum with p predictors $x^* \in R^p$ is given, SDA (Clemmensen, Hastie, and Ersbøll 2008) implemented nearest class mean rule to predict a group for x^* . It assigns x^* into category j_0, such taht

$$j_0 = \arg\min_j (x^{*\prime}\hat{\boldsymbol{\beta}} - M_j'\hat{\boldsymbol{\beta}})'(\Sigma_W + \lambda_2\boldsymbol{\Omega})(x^{*\prime}\hat{\boldsymbol{\beta}} - M_j'\hat{\boldsymbol{\beta}}), \tag{2}$$

where M_j is the mean of predictors in the jth category, Σ_W is the so-called within common covariance matrix and $\boldsymbol{\Omega}$ is the penalty matrix.

Prediction of Multi-label

In multi-label learning settings, the above nearest class mean rule cannot be applied because each data example belongs to multiple categories rather than exactly one category. From Equation (2), we see that each point (x, y) is represented by $x'\boldsymbol{\beta}$ and nearest class mean rule is implemented on the transformed point. By using optimal scoring and solving optimization problem (1), $x'\boldsymbol{\beta}$ can be approximated by $y\boldsymbol{\theta}$ with learned $\boldsymbol{\theta}$. Therefore, we have two approximate ways to represent x , namely $x'\boldsymbol{\beta}$ and $y\boldsymbol{\theta}$.

Given new x^* with unknown class label y^*, we also can represent x^* with $x^{*\prime}\boldsymbol{\beta}$. Another potential representation $y^*\boldsymbol{\theta}$ of x^* can be obtained by minimizing the Euclidean distance between $x^{*\prime}\boldsymbol{\beta}$ and $y^*\boldsymbol{\theta}$ over all possible values of y^* and the label vector of x^* is the minimizer of the minimization problem.

Denote by $\mathbf{1}$ a vector with all elements equal to 1. For non-overlapping categories, y^* can be obtained by the following quadratic optimization problem:

$$\begin{aligned} \hat{y}^* = \arg \min_{y\in\{0,1\}^J} \|y\boldsymbol{\theta} - x^{*\prime}\hat{\boldsymbol{\beta}}\|_2 \\ s.t. \qquad y\mathbf{1} = 1. \end{aligned} \tag{3}$$

Suppose the estimated $\hat{y}^*$ has j_0th element 1 and 0 elsewhere, then x^* is classified to the j_0th class.

Let $\boldsymbol{\theta}_{j,:}$ be the jth row of $\boldsymbol{\theta}$. For the non-overlapping category case, when (x, y) belongs to the jth category, $y\boldsymbol{\theta} = \boldsymbol{\theta}_{j,:}$. So Equation (3) is equivalent to

$$j_0 = \arg\min_j \|\hat{\boldsymbol{\theta}}_{j,:} - x^{*\prime}\hat{\boldsymbol{\beta}}\|_2. \tag{4}$$

The equivalence is in the sense that the optimizer $\hat{y}$ of (3) has its j_0th element $\hat{y}_{j_0} = 1$ and 0 elsewhere. (4) has a very good geometric interpretation: each data example in the training data set is represented by $Y\boldsymbol{\theta}$ and all of the training data are distributed into J points, i.e., $\boldsymbol{\theta}_{j,:}$ $(j = 1, ..., J)$. The new unlabeled x^* is represented by $x^{*\prime}\boldsymbol{\beta}$ which is an approximation of $y^*\boldsymbol{\theta}$. x^* is assigned to the class whose centroid in the new represented space is the closest to $x^{*\prime}\boldsymbol{\beta}$ in term of Euclidean distance. From this geometric point of view, (4) is a nearest class mean rule.

For situations where the unlabeled x^* belongs to multiple categories, we cannot apply nearest class mean rules such as (2) and (4) anymore. The formula (3) can be extended to (5) by removing the constraint that each data example only belongs to one category. The prediction of y^* for x^* with multiple labels is then:

$$\hat{y}^* = \arg \min_{y\in\{0,1\}^J} \|y\boldsymbol{\theta} - x^{*\prime}\hat{\boldsymbol{\beta}}\|_2. \tag{5}$$

The only difference between (3) and (5) is that the latter equation does not have the single-category constraint, while the former one has this constraint.

Choice of Penalty Matrix Ω

The quadratic penalty is used to avoid overfitting when predictors have high correlations or the number of predictors are much larger than the number of data samples ($p \gg n$). Several different penalty matrices have been proposed for various practical applications. For example, a second derivative-type penalty matrix is used for hyperspectral data analysis (Yu et al. 1999). In (Clemmensen, Hastie, and Ersbøll 2008), the identity matrix $\boldsymbol{\Omega} = \mathbf{I}_{p\times p}$ is used. But there are no theoretical results about how to choose a good penalty matrix. In this paper, we consider a more general penalty matrix than the identity matrix. We choose $\boldsymbol{\Omega}$ to be an equicorrelation matrix $\boldsymbol{\Delta}$, defined as follows.

Definition 1 (Equicorrelation Matrix). *Equicorrelation matrix $\boldsymbol{\Delta}$ is a symmetric definite matrix,*

$$\begin{aligned} \boldsymbol{\Delta} &= (1-\rho)\mathbf{I}_{p\times p} + \rho \mathbf{1}_p \mathbf{1}_p' \\ &= \begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}, (1-p)^{-1} < \rho < 1, \end{aligned}$$

where $\mathbf{1}_p$ is a $p \times 1$ vector with all of its elements 1.

This equicorrelation matrix is more general than the identity matrix, because when $\rho = 0$, it specializes to the identity matrix.

We argue that SDA with equicorrelation matrix $\boldsymbol{\Delta}$ in formula (1) will not only avoid overfitting, but also produce a grouping effect. That is, given the jth task, the distance

of the coefficients of highly positively correlated predictors will tend to be close when they have the same sign, if λ_2 is large and ρ is not too close to 1.

Theorem 1. *Assume that predictors $\mathbf{X}_l(l = 1, \ldots, p)$, are normalized such that $\frac{1}{n}\sum_{i=1}^{n} X_{il} = 0$ and $\frac{1}{n}\sum_{i=1}^{n} X_{il}^2 = 1$. Let $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}$ denote the solution of* (1)*, and $r_{l_1 l_2}$ the empirical correlation of $\mathbf{X}_{l_1}$ and $\mathbf{X}_{l_2}$. Define the distance of l_1th predictor and l_2th predictor for jth task as*

$$D_{l_1,l_2}^{(j)} = \frac{\left|\hat{\boldsymbol{\beta}}_{l_1 j} - \hat{\boldsymbol{\beta}}_{l_2 j}\right|}{\|\mathbf{Y}\hat{\boldsymbol{\theta}}_j\|_2}.$$

If $sign(\hat{\boldsymbol{\beta}}_{l_1 j}) = sign(\hat{\boldsymbol{\beta}}_{l_2 j})$ then we have

$$D_{l_1,l_2}^{(j)} \leq \frac{\sqrt{2(1 - r_{l_1 l_2})}}{\lambda_2(1-\rho)}.$$

A proof of Theorem 1 can be found in the appendix.

Algorithm of MtSDA

This section provides detailed descriptions on the estimation of parameters in (1) and the prediction of class label for unlabeled data.

Parameter Estimation There are two parameter vectors to be estimated, i.e., $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ in (1). We design an iterative optimization algorithm for (1).

For fixed $\boldsymbol{\theta}$ in (1) we obtain the *elastic-net* like problem for the jth $(j = 1, \ldots, J)$ task as follows

$$\hat{\boldsymbol{\beta}}_j = \arg\min_{\boldsymbol{\beta}_j} \frac{1}{n}\left(||\mathbf{Y}\boldsymbol{\theta}_j - \mathbf{X}\boldsymbol{\beta}_j||_2^2 + \lambda_2 \boldsymbol{\beta}_j' \boldsymbol{\Delta} \boldsymbol{\beta}_j + \lambda_1 ||\boldsymbol{\beta}_j||_1\right). \tag{6}$$

In this section, we assume ρ is known. When it is not as in the data section, we use cross validation to choose one from data. Equation (6) takes the form of a modified *naïve elastic net* problem (Zou and Hastie 2005). Since the matrix $\boldsymbol{\Delta}$ is symmetric and positive definite, $\sqrt{\boldsymbol{\Delta}}$ always exists. Define an artificial data set $(\tilde{\mathbf{y}}, \tilde{\mathbf{X}})$ by

$$\tilde{\mathbf{X}}_{(n+p)\times p} = (1+\lambda_2)^{-1/2}\begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\boldsymbol{\Delta} \end{pmatrix}, \tilde{\mathbf{y}}_{(n+p)} = \begin{pmatrix} \mathbf{Y}\boldsymbol{\theta}_j \\ \mathbf{0} \end{pmatrix}.$$

Let $\gamma = \lambda_1/\sqrt{(1+\lambda_2)}$ and $\tilde{\beta} = \sqrt{(1+\lambda_2)}\beta$, then (6) can be written as

$$\mathcal{L}(\gamma, \tilde{\beta}) = ||\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\tilde{\beta}||_2^2 + \gamma||\tilde{\beta}||_1,$$

which is a standard *lasso* problem (Tibshirani 1996). Performing the LARS algorithm (Tibshirani 1996) on this augmented problem yields the solution of (6).

For fixed $\boldsymbol{\beta}$ we have

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \frac{1}{n}||\mathbf{Y}\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta}||_2^2. \\ s.t. \quad & \frac{1}{n}(\mathbf{Y}\boldsymbol{\theta})'(\mathbf{Y}\boldsymbol{\theta}) = I. \end{aligned} \tag{7}$$

Rewrite (7) as

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} ||n^{-1/2}\mathbf{Y}\boldsymbol{\theta} - n^{-1/2}\mathbf{X}\boldsymbol{\beta}||_2^2, \\ s.t. \quad & \left(n^{-1/2}\mathbf{Y}\boldsymbol{\theta}\right)'\left(n^{-1/2}\mathbf{Y}\boldsymbol{\theta}\right) = I, \end{aligned} \tag{8}$$

Algorithm 1 The Calculation of Threshold

Input training data $X \in R^{n\times p}$ as well as its corresponding indicator matrix $Y \in \{0,1\}^{n\times J}$ and predicted class label $\hat{Y} \in R^{n\times J}$ by (11).
1: Convert $\hat{\mathbf{Y}} \in R^{n\times J}$ into $1 \times (n \times J)$ vector S.
2: Sort S and set the values of the first k_0 ($k_0 = 1, \ldots, n$) smallest values of $\hat{Y}$ as 0 and the rest values of $\hat{Y}$ as 1, calculate the F-measure.
3: Obtain the final k_0 which achieve best F-measure.
4: Set the threshold $thresh$ as $(S_{k_0} + S_{k_0+1})/2$.
Output threshold.

which is a standard Procrustes problem (Eldén and Park 1999). Taking $\mathbf{SVD}\left(n^{-1/2}\mathbf{X}\boldsymbol{\beta}\right) = \mathbf{USV}'$, by solution of Procrustes we have $n^{-1/2}\mathbf{Y}\boldsymbol{\theta} = \mathbf{UV}'$. Thus the solution for (7) is

$$\hat{\boldsymbol{\theta}} = n^{1/2}\mathbf{Y}^{\dagger}\mathbf{UV}', \tag{9}$$

where $\mathbf{Y}^{\dagger}$ denotes the Moore-Penrose inverse (Golub and Van Loan 1996) of $\mathbf{Y}$.

Prediction of Labels After $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ are estimated, the label matrix $\hat{\mathbf{Y}}^*$ for m test examples $\mathbf{X}^* \in R^{m\times p}$ is given by (5) discussed before . That is,

$$\hat{\mathbf{Y}}^* = \arg\min_{\mathbf{Y}} \|\mathbf{Y}\boldsymbol{\theta} - \mathbf{X}^*\boldsymbol{\beta}\|_2^2, \tag{10}$$

the solution of which is

$$\hat{\mathbf{Y}}^* = \mathbf{X}^*\boldsymbol{\beta}\boldsymbol{\theta}^{\dagger}. \tag{11}$$

It is apparent that the values in $\hat{\mathbf{Y}}^*$ are continuous rather than binary. Here a threshold is learned to quantize $\hat{\mathbf{Y}}^*$ by algorithm 1. After the $thresh$ is learned from training data, we could use $thresh$ to quantize $\hat{\mathbf{Y}}^*$.

We summarize our MtSDA in Algorithm 2.

Algorithm 2 MtSDA for Classification with Overlapping Categories

Input training data matrix $\mathbf{X} \in R^{n\times p}$, corresponding indicator matrix $\mathbf{Y} \in \{0,1\}^{n\times J}$ and test data matrix $\mathbf{X}^* \in R^{m\times p}$.
1: Initialize $\boldsymbol{\theta} = n^{1/2}\mathbf{Y}^{\dagger}\mathbf{E}_{:,1:J}$, where $\mathbf{E} = \mathbf{I}_{n\times n}$
2: For $l = 1, \ldots, J$ solve the *elastic-net* like problem (6);
3: For fixed $\boldsymbol{\beta}$ compute the optimal scores by (9);
4: Repeat step 2 and 3 until convergence;
5: Update $\boldsymbol{\beta}$ for fixed $\boldsymbol{\theta}$ by solving (6);
6: Compute $\hat{\mathbf{Y}}^*$ by (11) and Algorithm 1.
Output indicator matrix $\hat{\mathbf{Y}}^*$ for test data.

Related Work

A straightforward approach to perform multi-label learning is to construct a binary classifier for each label. Instances relevant to each given label form the positive class, and the rest form the negative class (Joachims, Nedellec, and Rouveirol 1998). However, the one-against-the-rest approach

fails to keep the correlation information among different labels and significantly deteriorates the classification performance. Bucak *et al.* (2009) proposed multi-label ranking to address the multi-label learning problem. Multi-label ranking avoided constructing binary classifiers and intended to order all the relevant classes at higher rank than the irrelevant ones. Although multi-label ranking is applicable when the number of classes is very large, it suffers from the inability of capturing the correlation of labels. Recently, a number of approaches have been developed for multi-label learning that exploit the correlation among labels such as $RankSVM$ (Elisseeff and Weston 2002) and Hypergraph learning (Sun, Ji, and Ye 2008). When structure can be imposed on the label space, some *sparsity*-based multi-task learning approaches were used with a mixed $(2,1)$-norm (Argyriou, Evgeniou, and Pontil 2008) to induce common features across tasks. In this paper, the structure of the prediction coefficients is also sparse, but we do not have any other structure information on them. Actually, our data does not seem to support further structure information.

Experiments

We use data from three open benchmark data collections to evaluate our method. The first one is the 11 top-level multi-topic webpage data set from Yahoo (Kazawa et al. 2005; Ji et al. 2008). For details of the Yahoo data collection please refer to Ji et al. (2008). The second is annotated images from the NUS-WIDE data collection (Chua et al. 2009) with multiple tags. We randomly sampled $10,000$ images from the NUS-WIDE data collection. Five types of low-level visual features (predictors) extracted from these images were concatenated and normalized into one 634-dimension predictor vector for each image. For the ground truth of image annotation we chose two indicator matrices for the selected data samples to form two data sets– NUS-6 and NUS-16. For the two data sets, the top 6 and 16 tags which label the maximum numbers of positive instances (NPI) were respectively selected. So, the correlations between tags in NUS-6 is more dense than those in NUS-16. The third annotated image data set is from the MSRA-MM dataset (Version 2.0) (Wang, Yang, and Hua 2009). We randomly sampled 10,000 multi-tagged images and the top 25 tags which label the maximum number of positive instances (NPI) were selected. Six types of low-level visual features were extracted from these images and concatenated and normalized into one 892-dimension vector for each image. We call this dataset MM2.0. We summarize some statistics for these data sets in Table 1. MaxNPI and MinNPI denote the maximum and the minimum number of positive instances for each topic (label) respectively in Table 1.

For the training data, we randomly sampled 1,000 samples from each of the 11 Yahoo data sets same as (Ji et al. 2008). For each of NUS-6, NUS-16 and MM2.0, we randomly sampled 5,000 samples as training data. The remaining data were used as the corresponding test data. This process was repeated five times to generate five random training/test partitions. At the first random partition, we tuned the tuning parameters ρ, λ_1 and λ_2. These tuning parameters were then fixed to train the MtSDA model for all these five partitions. For each partition, we obtained the prediction performance. The average performance and standard deviations are reported.

Table 1: Statistics for the data sets used. J, p, and N denote the number of tasks, the number of predictors, and the total number of instances. MaxNPI and MinNPI denote the maximum and the minimum number of positive instances for each topic (label) respectively. The statistics of the Yahoo data collection are the average of the 11 data sets.

Dataset	J	p	N	MaxNPI	MinNPI
Yahoo	18	23,970	10,626	4,488	129
NUS-6	6	634	10,000	2,686	1,004
NUS-16	16	634	10,000	2,686	337
MM2.0	25	892	10,000	3,988	107

Different penalty matrices can be used in MtSDA for multi-label learning. The MtSDA with identity matrix $\mathbf{I}$ and equicorrelation matrix $\mathbf{\Delta}$ are named as MtSDA ($\mathbf{\Omega} = \mathbf{I}$) and MtSDA ($\mathbf{\Omega} = \mathbf{\Delta}$) respectively. To evaluate classification and annotation performance, we compare MtSDA with two single-task learning algorithms and one recently developed multi-task learning algorithm. The single-task learning algorithms consist of SVM with a linear kernel and LDA+SVM. The SVM is applied on each label using a one-against-the-rest scheme. For LDA+SVM, the Linear Discriminant Analysis (LDA) is applied first on each label for dimension reduction before linear SVM is applied. The compared multi-task learning algorithm is HSML (Hypergraph spectral multi-label) (Sun, Ji, and Ye 2008). HSML constructs a hypergraph to capture the correlation information in different labels, and learns a lower-dimensional embedding in which the linear Support Vector Machine (SVM) is applied for each label separately.

The area under the ROC curve (AUC) is used to measure the performance for Webpage classification and image annotation. To measure the global performance across multiple tasks, according to (Lewis 1991) we use the microaveraging method. In order to measure the performance of image annotation, the F-measure (harmonic mean of precision and recall) is preferred.

The AUC scores from each algorithm are reported in Table 2. As a whole, MtSDA seems to provide much bet-

Table 3: Summary of performance for the four compared algorithms on NUS-6 , NUS-16 and MM2.0 in terms of F-measure. All parameters of the four algorithms are tuned by 5-fold cross-validation, and the average F-measure over 5 random sampling of training instances are reported. The highest performance is highlighted in each case.

Dataset	NUS-6	NUS-16	MM2.0
SVM	0.4018±0.0269	0.2207±0.0229	0.2089±0.0143
LDA+SVM	0.4018±0.0269	0.2207±0.0229	0.2089±0.0143
HSML	0.3994±0.0412	0.2463±0.0116	0.3626±0.0133
MtSDA ($\mathbf{\Omega}$=$\mathbf{I}$)	0.4232±0.0706	0.2945±0.0681	0.3509±0.0198
MtSDA ($\mathbf{\Omega}$=$\mathbf{\Delta}$)	**0.4554±0.0266**	**0.3410±0.0113**	**0.3629±0.0461**

Table 2: Summary of performance for the compared algorithms on 11 Yahoo data sets, NUS-6, NUS-16 and MM2.0 in terms of AUC. The average performance and standard deviations over 5 random training/test partitions are reported. The highest performance is highlighted in each case.

Dataset	SVM	LDA+SVM	HSML	MtSDA ($\mathbf{\Omega}$=$\mathbf{I}$)	MtSDA ($\mathbf{\Omega}$=$\mathbf{\Delta}$)
Arts	0.5357±0.0167	0.5238±0.0011	0.5771±0.0022	0.7610±0.0131	**0.7620±0.0119**
Business	0.8862±0.0070	0.8807±0.0005	0.8895±0.0011	0.8934±0.0171	**0.9034±0.0070**
Computers	0.7532±0.0030	0.7540±0.0008	0.7664±0.0014	0.8082±0.0152	**0.8184±0.0109**
Education	0.5915±0.0485	0.5645±0.0659	0.5922±0.0058	0.7200±0.1058	**0.7695±0.0135**
Entertainment	0.6605±0.0194	0.6137±0.0008	0.6608±0.0041	**0.8005±0.0055**	0.7978±0.0046
Health	0.7120±0.0241	0.6657±0.0059	0.7220±0.0052	0.8255±0.0577	**0.8494±0.0045**
Recreation	0.5789±0.0134	0.5589±0.0006	0.5967±0.0023	0.7699±0.0182	**0.7705±0.0216**
Reference	0.6927±0.0342	0.6453±0.0068	0.7041±0.0030	0.8267±0.0131	**0.8383±0.0045**
Science	0.5233±0.0067	0.5150±0.0060	0.5575±0.0056	0.7944±0.0178	**0.8035±0.0103**
Social	0.7463±0.0253	0.6268±0.0080	0.7471±0.0039	0.7925±0.1272	**0.8665±0.0282**
Society	0.6108±0.0802	0.6194±0.0538	0.6278±0.0051	0.7598±0.0174	**0.7604±0.0181**
NUS-6	0.6247±0.0777	0.6247±0.0777	0.6548±0.0709	0.6824±0.0316	**0.7453±0.0207**
NUS-16	0.6002±0.1123	0.6002±0.1123	0.6414±0.0775	0.5752±0.0716	**0.6533±0.0384**
MM2.0	0.7502±0.0182	0.7502±0.0182	**0.7726±0.0175**	0.7331±0.0415	0.7591±0.0240

ter multi-label classification performance than the other algorithms even in the $p \gg n$ situation, i.e., on average $p \approx 23,970$ and $n = 1,000$ for the 11 Yahoo data sets. The F-measure of image annotation is reported in Table 3. The MtSDA provides better annotation performance on the NUS-6, NUS-16 and MM2.0 data sets. Moreover, MtSDA ($\mathbf{\Omega} = \mathbf{\Delta}$) outperforms MtSDA ($\mathbf{\Omega} = \mathbf{I}$) slightly.

Figure shows the effect of the parameter ρ on MtSDA ($\mathbf{\Omega} = \mathbf{\Delta}$). It depicts the AUC scores of MSDA ($\mathbf{\Omega} = \mathbf{\Delta}$) on four data sets with respect to different value of the parameter $\rho = \frac{1}{k}$ on different $k \in \{1, 2, \ldots, 10\}$. When we changed ρ, the other tuning parameters λ_1 and λ_2 were fixed as the tuned values at the first random partition. We used the results of MtSDA ($\mathbf{\Omega} = \mathbf{I}$) as baseline. For each ρ, the average of AUC as well as its 95% confidence bounds over five random training/test partitions are plotted. 95% confidence bound of the average of a set $V = \{v_1, v_2, \ldots, v_s\}$ is defined as $\bar{V} \pm \frac{1.96\sigma_V}{\sqrt{s}}$, where $\bar{V}$ is the average of V and σ_V is the standard deviation of V. We can see that the parameter ρ generally has impact on the AUC of MtSDA ($\mathbf{\Omega} = \mathbf{\Delta}$). However, when the variability is taken into account, the impact of $\rho \neq 0$ is less obvious.

Conclusion

We have extended SDA to MtSDA for multi-label classification. A class of equicorrelation matrices is used in MtSDA which includes the identity matrix. Experiments show that the MtSDA achieves better results than single-task learning methods such as SVM and LDA+SVM. It also outperforms a multi-task learning method HSML in most cases.

Appendix

Proof of Theorem 1.

Proof. By setting the differentiation of objective function in formula (1) at $\hat{\boldsymbol{\beta}}_{l_1 j}$ and $\hat{\boldsymbol{\beta}}_{l_2 j}$ to zero (Boyd and Vandenberghe 2004), we have

$$-2\mathbf{X}_{l_1}^T(\mathbf{Y}\hat{\boldsymbol{\theta}}_j - \mathbf{X}\hat{\boldsymbol{\beta}}_j) + 2\lambda_2(\hat{\boldsymbol{\beta}}_{l_1 j} + \rho\sum_{l\neq l_1}\hat{\boldsymbol{\beta}}_l) + \lambda_1 sign(\hat{\boldsymbol{\beta}}_{l_1 j}) = 0, \tag{12}$$

and

$$-2\mathbf{X}_{l_2}^T(\mathbf{Y}\hat{\boldsymbol{\theta}}_j - \mathbf{X}\hat{\boldsymbol{\beta}}_j) + 2\lambda_2(\hat{\boldsymbol{\beta}}_{l_2 j} + \rho\sum_{l\neq l_2}\hat{\boldsymbol{\beta}}_l) + \lambda_1 sign(\hat{\boldsymbol{\beta}}_{l_2 j}) = 0, \tag{13}$$

where $sign(\hat{\boldsymbol{\beta}}_{ij})$ = 1 if $\hat{\boldsymbol{\beta}}_{ij} > 0$; $sign(\hat{\boldsymbol{\beta}}_{ij})$ = 0 if $\hat{\boldsymbol{\beta}}_{ij} = 0$; $sign(\hat{\boldsymbol{\beta}}_{ij})$ = -1 if $\hat{\boldsymbol{\beta}}_{ij} < 0$.

If $sign(\hat{\boldsymbol{\beta}}_{l_1 j}) = sign(\hat{\boldsymbol{\beta}}_{l_2 j})$, from the difference of the two equations (12) and (13), we have

$$-(2\mathbf{X}_{l_1} - 2\mathbf{X}_{l_2})'(\mathbf{Y}\hat{\boldsymbol{\theta}}_j - \mathbf{X}\hat{\boldsymbol{\beta}}_j) + 2\lambda_2(1-\rho)(\hat{\boldsymbol{\beta}}_{l_1 j} - \hat{\boldsymbol{\beta}}_{l_2 j}) = 0.$$

Then we have

$$\hat{\boldsymbol{\beta}}_{l_1 j} - \hat{\boldsymbol{\beta}}_{l_2 j} = \frac{(\mathbf{X}_{l_1} - \mathbf{X}_{l_2})'(\mathbf{Y}\hat{\boldsymbol{\theta}}_j - X\hat{\boldsymbol{\beta}}_j)}{\lambda_2(1-\rho)},$$

and by the definition of $D^{(j)}_{l_1,l_2}$

$$D^{(j)}_{l_1,l_2} = \frac{(\mathbf{X}_{l_1} - \mathbf{X}_{l_2})'(\mathbf{Y}\hat{\boldsymbol{\theta}}_j - X\hat{\boldsymbol{\beta}}_j)}{\lambda_2(1-\rho)\|\mathbf{Y}\hat{\boldsymbol{\theta}}_j\|_2}.$$

Because $\hat{\boldsymbol{\beta}}_j$ is the solution of optimization problem (6) when $\boldsymbol{\theta}_j = \hat{\boldsymbol{\theta}}_j$ is fixed, and that $\|Y\hat{\boldsymbol{\theta}}_j\|_2^2$ is the value of the objective function of (6) at the point $\beta_j = 0$, so we have that

$$\|\mathbf{Y}\hat{\boldsymbol{\theta}}_j - \mathbf{X}\hat{\boldsymbol{\beta}}_j\|_2^2 \leq \|\mathbf{Y}\hat{\boldsymbol{\theta}}_j\|_2^2.$$

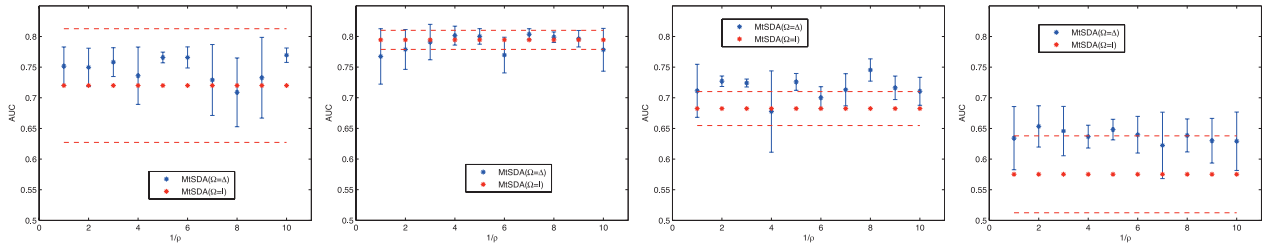
Because of the normalization of X,

$$\|\mathbf{X}_{l_1} - \mathbf{X}_{l_2}\|_2^2 = 2 - 2r_{l_1 l_2}.$$

So, by *Cauchy-Schwarz* inequality, we have

$$D^{(j)}_{l_1,l_2} \leq \frac{\sqrt{2(1-r_{l_1 l_2})}}{\lambda_2(1-\rho)}.$$

□



(a) Education (b) Science (c) NUS-6 (d) NUS-16

Figure 1: The effect of parameter ρ on AUC for four datasets from Education and Science categories of Yahoo Webpage as well as NUS-6 and NUS-16 when $\frac{1}{\rho}$ varies on $\{1, 2, \ldots, 10\}$. For each ρ, the average of AUC as well as its 95% confidence bounds over five random training/test partitions are plotted.

Acknowledgements

This work is supported by NSFC (No.90920303), 973 Program (No.2009CB320801). Fei Wu thanks the department of statistics for the invitation of his visiting sponsored by "New Star" plan from Zhejiang University and the work is done when he is visiting UC Berkeley. Jinzhu Jia is supported by the national science foundation, CDI grant SES-083553. Prof. Bin Yu is partially supported by the national science foundation, CDI grant SES-083553, DMS-0907632 and a grant from MSRA.

References

Argyriou, A.; Evgeniou, T.; and Pontil, M. 2008. Convex multi-task feature learning. *Machine Learning* 73(3):243–272.

Boyd, S., and Vandenberghe, L. 2004. *Convex Optimization*. Cambridge Univ Pr.

Bucak, S.; Mallapragada, P.; Jin, R.; and Jain, A. 2009. Efficient multi-label ranking for multi-class learning: application to object recognition. In *Proceedings of 12th IEEE International Conference on Computer Vision*.

Chua, T.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. Nus-wide: A real-world web image database from national university of singapore. In *ACM International Conference on Image and Video Retrieval*.

Clemmensen, L.; Hastie, T.; and Ersbøll, B. 2008. Sparse Discriminant Analysis. *online: http://www-stat.stanford.edu/ hastie/Papers/*.

Eldén, L., and Park, H. 1999. A Procrustes problem on the stiefel manifold. *Numerische Mathematik* 82(4):599–619.

Elisseeff, A., and Weston, J. 2002. Kernel methods for multi-labelled classification and categorical regression problems. In *Advances in Neural Information Processing Systems 14*.

Golub, G., and Van Loan, C. 1996. *Matrix computations*. Johns Hopkins Univ Pr.

Hastie, T.; Buja, A.; and Tibshirani, R. 1995. Penalized discriminant analysis. *The Annals of Statistics* 23(1):73–102.

Ji, S.; Tang, L.; Yu, S.; and Ye, J. 2008. Extracting shared subspace for multi-label classification. In *Proceeding of the 14th ACM International Conference on Knowledge Discovery and Data Mining*, 381–389.

Joachims, T.; Nedellec, C.; and Rouveirol, C. 1998. Text categorization with support vector machines: learning with many relevant. In *Machine Learning: ECML-98 10th European Conference on Machine Learning*, 137–142.

Kazawa, H.; Izumitani, T.; Taira, H.; and Maeda, E. 2005. Maximal margin labeling for multi-topic text categorization. *Advances in Neural Information Processing Systems 17* 649–656.

Lewis, D. 1991. Evaluating text categorization. In *Proceedings of Speech and Natural Language Workshop*, 312–318.

Lounici, K.; Tsybakov, A.; Pontil, M.; and van de Geer, S. 2009. Taking advantage of sparsity in multi-task learning. In *Proceedings of the Conference on Learning Theory*.

Sun, L.; Ji, S.; and Ye, J. 2008. Hypergraph spectral learning for multi-label classification. In *Proceeding of the 14th ACM International Conference on Knowledge Discovery and Data Mining*, 668–676.

Tibshirani, R. 1996. Regression shrinkage and selection via the *lasso*. *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1):267–288.

Wang, M.; Yang, L.; and Hua, X. 2009. MSRA-MM: Bridging research and industrial societies for multimedia information retrieval. *Microsoft Technical Report (MSR-TR-2009-30)*.

Yu, B.; Ostland, I.; Gong, P.; and Pu, R. 1999. Penalized discriminant analysis of in situ hyperspectral data for conifer species recognition. *IEEE Transactions on Geoscience and Remote Sensing* 37(5):2569–2577.

Yu, K.; Tresp, V.; and Schwaighofer, A. 2005. Learning Gaussian processes from multiple tasks. In *Proceedings of the 22nd International Conference on Machine Learning*, 1012–1019.

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B(Statistical Methodology)* 67(2):301–320.