

A Shared-Subspace Learning Framework for Multi-Label Classification

SHUIWANG JI and LEI TANG

Arizona State University

SHIPENG YU

Siemens Medical Solutions

and

JIEPING YE

Arizona State University

8

Multi-label problems arise in various domains such as multi-topic document categorization, protein function prediction, and automatic image annotation. One natural way to deal with such problems is to construct a binary classifier for each label, resulting in a set of independent binary classification problems. Since multiple labels share the same input space, and the semantics conveyed by different labels are usually correlated, it is essential to exploit the correlation information contained in different labels. In this paper, we consider a general framework for extracting shared structures in multi-label classification. In this framework, a common subspace is assumed to be shared among multiple labels. We show that the optimal solution to the proposed formulation can be obtained by solving a generalized eigenvalue problem, though the problem is nonconvex. For high-dimensional problems, direct computation of the solution is expensive, and we develop an efficient algorithm for this case. One appealing feature of the proposed framework is that it includes several well-known algorithms as special cases, thus elucidating their intrinsic relationships. We further show that the proposed framework can be extended to the kernel-induced feature space. We have conducted extensive experiments on multi-topic web page categorization and automatic gene expression pattern image annotation tasks, and results demonstrate the effectiveness of the proposed formulation in comparison with several representative algorithms.

This work was supported by NSF IIS-0612069, IIS-0812551, CCF-0811790, NIH R01-HG002516, and NGA HM1582-08-1-0016.

Authors' addresses: S. Ji, School of Computing, Informatics and Decision Systems Engineering and the Center for Evolutionary Medicine and Informatics of the Biodesign Institute, Arizona State University, Tempe, AZ 85287; email: shuiwang.ji@asu.edu; L. Tang, School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, AZ 85287; email: L.Tang@asu.edu; S. Yu, Computer-Aided Diagnosis Group, Siemens Medical Solutions, Malvern, PA 19355; email: shipeng.yu@siemens.com; J. Ye, School of Computing, Informatics and Decision Systems Engineering and the Center for Evolutionary Medicine and Informatics of the Biodesign Institute, Arizona State University, Tempe, AZ 85287; email: jieping.ye@asu.edu.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2010 ACM 1556-4681/2010/05-ART8 \$10.00
DOI 10.1145/1754428.1754431 <http://doi.acm.org/10.1145/1754428.1754431>

Categories and Subject Descriptors: H.2.8 [Database Management]: Database Applications—*Data mining*

General Terms: Algorithms

Additional Key Words and Phrases: Multi-label classification, kernel methods, shared subspace, least squares loss, singular value decomposition, web page categorization, gene expression pattern image annotation

ACM Reference Format:

Ji, S., Tang, L., Yu, S., and Ye, J. 2010. A shared-subspace learning framework for multi-label classification. *ACM Trans. Knowl. Discov. Data.* 4, 2, Article 8 (May 2010), 29 pages.

DOI = 10.1145/1754428.1754431 <http://doi.acm.org/10.1145/1754428.1754431>

1. INTRODUCTION

Learning from objects annotated with multiple labels is a frequently encountered and widely studied problem in many domains. For example, in web page categorization [Ueda and Saito 2002a; Kazawa et al. 2005; Ueda and Saito 2002b; Tang et al. 2009], a Web page can be assigned to multiple topics. In gene and protein function prediction [Barutcuoglu et al. 2006; Roth and Fischer 2007], multiple functional labels may be associated with each gene and protein, since an individual gene or protein usually performs multiple functions. In automatic image annotation [Barnard et al. 2003; Monay and Gatica-Perez 2007; Li and Wang 2008; Carneiro et al. 2007], multiple semantic labels are usually assigned to a single image to indicate the presence of multiple objects in it. One common aspect of these problems is that multiple labels are associated with a single object, and they are called multi-label classification problems. Such problems are more general than the traditional multi-class problems in which a single label is assigned to an object. Driven by various applications, such problems have recently received increasing attention [McCallum 1999; Jin and Ghahramani 2002; Elisseeff and Weston 2002; Yu et al. 2005; Zhang and Zhou 2006, 2007; Zhou and Zhang 2007; Sun et al. 2008a; Ghamrawi and McCallum 2005; Kang et al. 2006; Yan et al. 2007; Ji and Ye 2009].

One simple and popular approach for multi-label classification is to construct a binary classifier for each label in which instances relevant to this label form the positive class, and the rest form the negative class. This approach has been applied successfully to various applications [Fan and Lin 2007; Yang and Pedersen 1997; Joachims 1998]. However, it fails to capture the correlation information among different labels, which is critical for many applications where the semantics conveyed by different labels are correlated. Indeed, it has been shown that the decoupling of multiple labels may compromise the performance significantly in certain applications [Ueda and Saito 2002a]. For example, in modeling the topics and authorship of documents, it is evident that the topics and authors of documents are correlated, since a particular author may only write on certain topics. Hence, it is desirable to model them in a coordinated fashion so that their intrinsic relationships can be captured.

In this article, we propose a general framework for extracting shared structures (subspace) in multi-label classification. In this framework, a binary

classifier is constructed for each label to discriminate this label from the rest of them. However, unlike the approach that builds the binary classifiers independently, a low-dimensional subspace is assumed to be shared among multiple labels. The predictive functions in our formulation consist of two parts: the first part is contributed from the representations in the original data space, and the second one is contributed from the embedding in the shared subspace. A similar formulation has been proposed in Ando and Zhang [2005] for multi-task learning. We show that when the least squares loss is used in classification, the linear transformation that characterizes the shared subspace can be computed by solving a generalized eigenvalue problem. In contrast, the formulation proposed in Ando and Zhang [2005] is nonconvex and needs to be solved iteratively. For high-dimensional problems, direct computation of the solution is computationally expensive, and we develop an efficient algorithm for this case. One appealing feature of the proposed framework is that it includes several well-known algorithms as special cases, thus elucidating their intrinsic relationships. We further show that the proposed framework can be extended to the kernel-induced feature space. We have conducted extensive experiments on web page categorization and automatic gene expression pattern image annotation tasks, and results demonstrate the effectiveness of the proposed formulations. Experimental results also show that the proposed formulations based on the least squares loss is comparable to other formulations based on the hinge loss, while it is much more efficient.

The key contributions of this article are highlighted as follows.

- We propose a general framework for extracting shared structures in multi-label classification. In this framework, the correlation information among multiple labels is captured by a low-dimensional subspace shared among all labels.
- We show that when the least squares loss is used in classification, the shared structure can be computed by solving a generalized eigenvalue problem. To reduce the computational cost, we propose an efficient algorithm for high-dimensional problems.
- We show that the proposed formulation includes several well-known formulations as special cases and further extend it to the kernel-induced feature space.
- We have conducted extensive experiments on multi-topic web page categorization and automatic gene expression pattern image annotation tasks to demonstrate the effectiveness of the proposed formulation.

The rest of this article is organized as follows: We present the framework for extracting shared subspace in Section 2. The efficient algorithm for computing the solution is presented in Section 3. We discuss its relationship with existing formulations in Section 4. In Section 5, the proposed formulation is extended to the kernel-induced feature space. We report experimental results in Section 6 and conclude this article by discussing further research in Section 7.

Notations. We use n , d , and m to denote the number of training instances, the data dimensionality, and the number of labels, respectively. The data matrix

and the label indicator matrix are denoted as $X = [x_1, \dots, x_n]^T \in \mathbb{R}^{n \times d}$ and $Y \in \mathbb{R}^{n \times m}$, where $x_i \in \mathbb{R}^d$ is the i th instance, and $Y_{i\ell} = 1$ if the i th instance has the ℓ th label, and -1 otherwise.

2. THE PROPOSED FRAMEWORK

We are given a set of input data $\{x_i\}_{i=1}^n \in \mathbb{R}^d$ and the class label indicator matrix $Y \in \mathbb{R}^{n \times m}$ that encodes the label information, where m and n are the number of labels and the number of instances, respectively. Following the traditional supervised learning framework, we learn m functions $\{f_\ell\}_{\ell=1}^m$ from the data that minimize the following regularized empirical risk:

$$R(\{f_\ell\}_{\ell=1}^m) = \sum_{\ell=1}^m \left(\frac{1}{n} \sum_{i=1}^n L(f_\ell(x_i), y_i^\ell) + \mu \Omega(f_\ell) \right), \quad (1)$$

where $y_i^\ell = Y_{i\ell}$, L is a prescribed loss function, $\Omega(f)$ is a regularization functional measuring the smoothness of f , and $\mu > 0$ is the regularization parameter.

2.1 Problem Formulation

We propose a multi-label learning framework, in which a low-dimensional subspace is shared by all labels. The predictive functions in this framework consist of two parts: one part is contributed from the original data space, and the other part is derived from the shared subspace as follows:

$$f_\ell(x) = \mathbf{w}_\ell^T x + \mathbf{v}_\ell^T \Theta x, \quad (2)$$

where $\mathbf{w}_\ell \in \mathbb{R}^d$ and $\mathbf{v}_\ell \in \mathbb{R}^r$ are the weight vectors, $\Theta \in \mathbb{R}^{r \times d}$ is the linear transformation used to parameterize the shared low-dimensional subspace, and r is the dimensionality of the shared subspace. The transformation Θ is common for all labels, and it has orthonormal rows, that is $\Theta \Theta^T = I$. In this formulation, the input data are projected onto a low-dimensional subspace by Θ , and this low-dimensional projection is combined with the original representation to produce the final prediction. Note that a similar formulation has been proposed in Ando and Zhang [2005] to capture the shared predictive structures in multi-task learning, and our formulation differs from it in several key aspects (see Section 4.1 for a comparison).

Following the regularization formulation in Equation (1), we propose to estimate the parameters $\{\mathbf{w}_\ell, \mathbf{v}_\ell\}_{\ell=1}^m$ and Θ by minimizing the following regularized empirical risk:

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

subject to the constraint that $\Theta \Theta^T = I$. Note that in the above formulation, the first regularization term $\|\mathbf{w}_\ell\|^2$ controls the amount of information specific to each label, while the second regularization term $\|\mathbf{w}_\ell + \Theta^T \mathbf{v}_\ell\|^2$ controls the complexity of the models for each label. By a change of variable, this problem

can be reformulated equivalently as follows:

$$\begin{aligned} \min_{\{\mathbf{u}_\ell, \mathbf{v}_\ell\}, \Theta} \quad & \sum_{\ell=1}^m \left(\frac{1}{n} \sum_{i=1}^n L(\mathbf{u}_\ell^T x_i, y_i^\ell) + \alpha \|\mathbf{u}_\ell - \Theta^T \mathbf{v}_\ell\|^2 + \beta \|\mathbf{u}_\ell\|^2 \right) \\ \text{s. t.} \quad & \Theta \Theta^T = I. \end{aligned} \quad (3)$$

In this article, we consider the least squares loss, that is,

$$L(\mathbf{u}_\ell^T x_i, y_i^\ell) = (\mathbf{u}_\ell^T x_i - y_i^\ell)^2.$$

It has been shown [Fung and Mangasarian 2005; Rifkin and Klautau 2004] that the least squares loss function is comparable to other loss functions such as the hinge loss employed in support vector machines (SVM) [Schölkopf and Smola 2002] when appropriate regularization is added. Hence, we get the following optimization problem:

$$\begin{aligned} \min_{\{\mathbf{u}_\ell, \mathbf{v}_\ell\}, \Theta} \quad & \sum_{\ell=1}^m \left(\frac{1}{n} \|X \mathbf{u}_\ell - y_\ell\|^2 + \alpha \|\mathbf{u}_\ell - \Theta^T \mathbf{v}_\ell\|^2 + \beta \|\mathbf{u}_\ell\|^2 \right) \\ \text{s. t.} \quad & \Theta \Theta^T = I, \end{aligned} \quad (4)$$

where $X = [x_1, \dots, x_n]^T \in \mathbb{R}^{n \times d}$ is the data matrix, $y_\ell = [y_1^\ell, \dots, y_n^\ell]^T \in \mathbb{R}^n$. The formulation in Equation (4) can be expressed compactly as:

$$\begin{aligned} \min_{U, V, \Theta} \quad & \frac{1}{n} \|XU - Y\|_F^2 + \alpha \|U - \Theta^T V\|_F^2 + \beta \|U\|_F^2 \\ \text{s. t.} \quad & \Theta \Theta^T = I, \end{aligned} \quad (5)$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix [Golub and Van Loan 1996], $U = [\mathbf{u}_1, \dots, \mathbf{u}_m]$, and $V = [\mathbf{v}_1, \dots, \mathbf{v}_m]$.

2.2 The Computation of V^*

We show that the optimal V^* that solves the optimization problem in Equation (5) can be expressed in terms of Θ and U , as summarized in the following lemma:

LEMMA 2.1. *Let U , V , and Θ be defined as before. Then the optimal V^* that solves the optimization problem in Equation (5) is given by $V^* = \Theta U$.*

PROOF. The only term in Equation (5) that depends on V is $\|U - \Theta^T V\|_F^2$, which can be expressed equivalently as:

$$\begin{aligned} \|U - \Theta^T V\|_F^2 &= \text{tr}((U^T - V^T \Theta)(U - \Theta^T V)) \\ &= \text{tr}(U^T U + V^T \Theta \Theta^T V - 2U^T \Theta^T V), \end{aligned} \quad (6)$$

where $\text{tr}(\cdot)$ denotes the trace of a matrix, and we have used the property that

$$\|A\|_F^2 = \text{tr}(A^T A)$$

for any matrix A . Taking the derivative of the expression in Equation (6) with respect to V , and setting it to zero, we obtain

$$V^* = \Theta U,$$

where we have used the property that $\Theta\Theta^T = I$. This completes the proof of the lemma. \square

2.3 The Computation of U^*

It follows from Lemma 2.1 that the objective function in Equation (5) can be rewritten as:

$$\begin{aligned} & \frac{1}{n} \|XU - Y\|_F^2 + \alpha \|U - \Theta^T V\|_F^2 + \beta \|U\|_F^2 \\ &= \frac{1}{n} \|XU - Y\|_F^2 + \alpha \|U - \Theta^T \Theta U\|_F^2 + \beta \|U\|_F^2 \\ &= \frac{1}{n} \|XU - Y\|_F^2 + \text{tr}(U^T((\alpha + \beta)I - \alpha\Theta^T\Theta)U). \end{aligned} \quad (7)$$

Hence, the optimization problem in Equation (5) can be expressed equivalently as:

$$\begin{aligned} & \min_{U, \Theta} \frac{1}{n} \|XU - Y\|_F^2 + \text{tr}(U^T((\alpha + \beta)I - \alpha\Theta^T\Theta)U) \\ & \text{s. t. } \Theta\Theta^T = I. \end{aligned} \quad (8)$$

We show that the optimal U^* can be expressed in terms of Θ . This is summarized in the following lemma:

LEMMA 2.2. *Let X, Y, U , and Θ be defined as before. Then the optimal U^* that solves the optimization problem in Equation (8) can be expressed as:*

$$U^* = \frac{1}{n}(M - \alpha\Theta^T\Theta)^{-1}X^TY, \quad (9)$$

where M is defined as:

$$M = \frac{1}{n}X^TX + (\alpha + \beta)I. \quad (10)$$

PROOF. Taking the derivative of the objective function in Equation (8) with respect to U , and setting it to zero, we obtain

$$U^* = \frac{1}{n}(M - \alpha\Theta^T\Theta)^{-1}X^TY, \quad (11)$$

where M is defined in Equation (10). \square

2.4 The Computation of Θ^*

It follows from Lemma 2.2 that we can substitute the expression for U^* in Equation (9) into Equation (8) and obtain the following optimization problem with respect to Θ :

$$\begin{aligned} & \max_{\Theta} \frac{1}{n^2} \text{tr}(Y^T X (M - \alpha\Theta^T\Theta)^{-1} X^T Y) \\ & \text{s. t. } \Theta\Theta^T = I. \end{aligned} \quad (12)$$

We show in the following theorem that the optimization problem in Equation (12) can be simplified, and the optimal Θ^* can be obtained by solving a generalized eigenvalue problem.

THEOREM 2.3. *Let X , Y , and Θ be defined as before. Then the optimal Θ^* that solves the optimization problem in Equation (12) can be obtained by solving the following trace maximization problem:*

$$\begin{aligned} \max_{\Theta} \quad & \text{tr}((\Theta S_1 \Theta^T)^{-1} \Theta S_2 \Theta^T) \\ \text{s. t.} \quad & \Theta \Theta^T = I, \end{aligned} \quad (13)$$

where S_1 and S_2 are defined as:

$$S_1 = I - \alpha M^{-1}, \quad (14)$$

$$S_2 = M^{-1} X^T Y Y^T X M^{-1}, \quad (15)$$

and M is defined in Equation (10).

PROOF. We need the Sherman-Woodbury-Morrison formula [Golub and Van Loan 1996] for computing matrix inverse:

$$(P + ST)^{-1} = P^{-1} - P^{-1} S (I + T P^{-1} S)^{-1} T P^{-1}. \quad (16)$$

It follows from the formula in Equation (16) that

$$\begin{aligned} (M - \alpha \Theta^T \Theta)^{-1} &= M^{-1} + \alpha M^{-1} \Theta^T (I - \alpha \Theta M^{-1} \Theta^T)^{-1} \Theta M^{-1} \\ &= M^{-1} + \alpha M^{-1} \Theta^T (\Theta (I - \alpha M^{-1}) \Theta^T)^{-1} \Theta M^{-1}, \end{aligned} \quad (17)$$

where the last equality follows since $\Theta \Theta^T = I$. By substituting the expression in Equation (17) into the optimization problem in Equation (12), we obtain the following problem:

$$\begin{aligned} \max_{\Theta} \quad & \text{tr}(Y^T X M^{-1} \Theta^T (\Theta (I - \alpha M^{-1}) \Theta^T)^{-1} \Theta M^{-1} X^T Y) \\ \text{s. t.} \quad & \Theta \Theta^T = I, \end{aligned}$$

where we have omitted the term $Y^T X M^{-1} X^T Y$ since it is independent of Θ . By using the property that $\text{tr}(AB) = \text{tr}(BA)$ for any two matrices A and B , and noticing the definitions of S_1 and S_2 in Equations (14) and (15), respectively, we prove this theorem. \square

Let $Z = [z_1, \dots, z_r]$ be the matrix consisting of the top r eigenvectors corresponding to the largest r nonzero eigenvalues of the generalized eigenvalue problem: $S_1 z = \lambda S_2 z$. To ensure the constraint $\Theta \Theta^T = I$, the QR decomposition can be employed. In particular, let $Z = Z_q Z_r$ be the QR decomposition of Z , where Z_q has orthonormal columns and Z_r is upper triangular. It is easy to verify [Ye 2005] that the objective function in Equation (13) is invariant of any nonsingular transformation, that is, Q and NQ achieve the same objective value for any nonsingular matrix $N \in \mathbb{R}^{r \times r}$. It follows that the optimal Q^* solving Equation (13) is given by $Q^* = Z_q^T$. Note that S_1 is positive definite (see Equation (19)), thus Z can also be obtained by computing the top eigenvectors of $S_1^{-1} S_2$.

3. AN EFFICIENT ALGORITHM

From the discussions in the last section, the optimal Θ^* is given by the eigenvectors of $S_1^{-1}S_2 \in \mathbb{R}^{d \times d}$ corresponding to the r largest eigenvalues. When the data dimensionality, that is, d , is small, the eigenvectors of $S_1^{-1}S_2$ can be computed directly. However, when d is large, direct eigendecomposition is computationally expensive. In this section, we show how we can compute the eigenvectors efficiently for this case.

3.1 Reformulation of $S_1^{-1}S_2$

It follows from the Sherman-Woodbury-Morrison formula in Equation (16) that

$$\begin{aligned} M^{-1} &= \frac{1}{\alpha + \beta} I - \frac{1}{n(\alpha + \beta)^2} X^T \left(I + \frac{1}{n(\alpha + \beta)} XX^T \right)^{-1} X \\ &= \frac{1}{\alpha + \beta} I - \frac{1}{\alpha + \beta} X^T (n(\alpha + \beta)I + XX^T)^{-1} X. \end{aligned} \quad (18)$$

Hence, we have

$$I - \alpha M^{-1} = \frac{\beta}{\alpha + \beta} I + \frac{\alpha}{\alpha + \beta} X^T (n(\alpha + \beta)I + XX^T)^{-1} X, \quad (19)$$

which is positive definite when $\beta > 0$.

It follows from the definitions of M , S_1 , and S_2 in Equations (10) (14), and (15) that

$$\begin{aligned} S_1^{-1}S_2 &= (I - \alpha M^{-1})^{-1} M^{-1} X^T Y Y^T X M^{-1} \\ &= (M - \alpha I)^{-1} X^T Y Y^T X M^{-1} \\ &= \left(\frac{1}{n} X^T X + \beta I \right)^{-1} X^T Y Y^T X \left(\frac{1}{n} X^T X + (\alpha + \beta) I \right)^{-1}. \end{aligned} \quad (20)$$

Let

$$X = U \Sigma V^T \quad (21)$$

be the singular value decomposition (SVD) [Golub and Van Loan 1996] of X , where $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{d \times d}$ are orthogonal,

$$\Sigma = \text{diag}(\Sigma_t, 0) \in \mathbb{R}^{n \times d}$$

is diagonal, and $t = \text{rank}(X)$. Let $U = [U_1, U_2]$, where $U_1 \in \mathbb{R}^{n \times t}$, $U_2 \in \mathbb{R}^{n \times (n-t)}$, $V = [V_1, V_2]$, $V_1 \in \mathbb{R}^{d \times t}$, $V_2 \in \mathbb{R}^{d \times (d-t)}$, and Σ_t consists of the first t rows and the

first t columns of Σ . Then we have

$$\begin{aligned}
S_1^{-1}S_2 &= V_1 \left(\frac{1}{n}\Sigma_t^2 + \beta I \right)^{-1} V_1^T X^T Y Y^T X \left(\frac{1}{n}X^T X + (\alpha + \beta)I \right)^{-1} \\
&\quad + \frac{1}{\beta} V_2 I^{-1} V_2^T X^T Y Y^T X \left(\frac{1}{n}X^T X + (\alpha + \beta)I \right)^{-1} \\
&= V_1 \left(\frac{1}{n}\Sigma_t^2 + \beta I \right)^{-1} V_1^T X^T Y Y^T X \left(\frac{1}{n}X^T X + (\alpha + \beta)I \right)^{-1} \\
&= V_1 \left(\frac{1}{n}\Sigma_t^2 + \beta I \right)^{-1} V_1^T X^T Y Y^T X V_1 \left(\frac{1}{n}\Sigma_t^2 + (\alpha + \beta)I \right)^{-1} V_1^T \\
&= V_1 \left(\frac{1}{n}\Sigma_t^2 + \beta I \right)^{-1} \Sigma_t U_1^T Y Y^T U_1 \Sigma_t \left(\frac{1}{n}\Sigma_t^2 + (\alpha + \beta)I \right)^{-1} V_1^T.
\end{aligned}$$

The second and the third equalities follow since the columns of V_2 are in the null space of X , that is,

$$XV_2 = 0.$$

3.2 Diagonalization of $S_1^{-1}S_2$

Define three diagonal matrices D_1 , D_2 , and D as follows:

$$D_1 = \left(\frac{1}{n}\Sigma_t^2 + \beta I \right)^{-1} \Sigma_t \in \mathbb{R}^{t \times t}, \quad (22)$$

$$D_2 = \Sigma_t \left(\frac{1}{n}\Sigma_t^2 + (\alpha + \beta)I \right)^{-1} \in \mathbb{R}^{t \times t}, \quad (23)$$

$$D = (D_1 D_2^{-1})^{\frac{1}{2}} \in \mathbb{R}^{t \times t}. \quad (24)$$

Then we have

$$\begin{aligned}
S_1^{-1}S_2 &= V_1 D_1 U_1^T Y Y^T U_1 D_2 V_1^T \\
&= V_1 D (D^{-1} D_1) U_1^T Y Y^T U_1 (D_2 D) D^{-1} V_1^T \\
&= V_1 D \tilde{D} U_1^T Y Y^T U_1 \tilde{D} D^{-1} V_1^T,
\end{aligned}$$

where

$$\tilde{D} = D^{-1} D_1 = D_2 D. \quad (25)$$

Denote $C = Y^T U_1 \tilde{D} \in \mathbb{R}^{m \times t}$ and let

$$C = P_1 \Lambda P_2^T \quad (26)$$

be the SVD of C where $P_1 \in \mathbb{R}^{m \times m}$ and $P_2 \in \mathbb{R}^{t \times t}$ are orthogonal, and $\Lambda \in \mathbb{R}^{m \times t}$ is diagonal. Then

$$\begin{aligned}
S_1^{-1}S_2 &= V_1 D P_2 \Lambda^T \Lambda P_2^T D^{-1} V_1^T \\
&= V_1 D P_2 \tilde{\Lambda} P_2^T D^{-1} V_1^T,
\end{aligned} \quad (27)$$

where $\tilde{\Lambda} = \Lambda^T \Lambda \in \mathbb{R}^{t \times t}$.

Table I.
Summary of relevant matrices. The size, the computation required, and the associated complexity of each relevant matrix are Listed

Matrix	Size	Computation	Complexity
X	$n \times d$	SVD	$O(dn^2)$
C	$m \times t$	SVD	$O(tm^2)$
V_1DP_2	$d \times t$	QR	$O(dt^2)$

3.3 Algorithms for Computing Θ^* and U^*

We can observe that Equation (27) gives the eigen-decomposition of $S_1^{-1}S_2$ corresponding to nonzero eigenvalues. Hence, the eigenvectors of $S_1^{-1}S_2$ corresponding to nonzero eigenvalues are given by the columns of V_1DP_2 . The algorithm for computing the optimal Θ^* for high-dimensional data is summarized as follows:

- 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 in Equations. (22), (23), (24) and (25), respectively.
- 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_1DP_2 as $V_1DP_2 = QR$.
- The rows of the optimal Θ^* are given by the first r columns of the matrix Q .

After obtaining Θ^* , we need to compute the optimal U^* given by Equation (9). Note that the matrix $M \in \mathbb{R}^{d \times d}$ is involved in Equation (9), and hence it is expensive to compute U^* directly for high-dimensional data. More specifically, we need to make use of the expressions in Equations (17), (18), and (19) so that explicit formations of the matrices M and M^{-1} are avoided.

The SVD of X in the first step takes $O(dn^2)$ time assuming $d > n$. The size of C is $m \times t$ where m is the number of labels and $t = \text{rank}(X)$. Hence the SVD of C in the third step takes $O(tm^2)$ time assuming $t > m$. The QR decomposition in the fourth step takes $O(dt^2)$ time. Typically, m and t are both small. Thus, the cost of the proposed algorithm for computing Θ^* is dominated by the cost for computing the SVD of X . A summary of relevant matrices and their associated computational complexities is listed in Table I.

4. RELATIONSHIP TO EXISTING ALGORITHMS

In this section, we show that the proposed formulation includes several well-known algorithms as special cases. We begin by discussing related work.

4.1 Related Work

4.1.1 Dimensionality Reduction. Canonical correlation analysis (CCA) [Hotelling 1936] and partial least squares (PLS) [Wold 1966; Arenas-García et al. 2007] are classical techniques for modeling relations between sets of observed variables. They both compute low-dimensional embedding of sets of variables simultaneously. Their main difference is that CCA maximizes the correlations between variables in the embedded space, while PLS maximizes their covariances. One popular use of CCA and PLS is for supervised learning,

in which one set of variables are derived from the data and another set are derived from the class labels [Sun et al. 2008b, 2009]. In this setting, the data can be projected onto a lower-dimensional space directed by the label information. Such formulation is particularly appealing in the context of dimensionality reduction for multi-label data. When applied to multi-class problems, CCA reduces to the well-known linear discriminant analysis (LDA) formulation [Fukunaga 1990] in which a projection is obtained by maximizing the ratio of interclass distance to intraclass distance.

4.1.2 Multi-Task Learning. In Ando and Zhang [2005], a similar formulation has been proposed for multi-task learning. In this formulation, the input data for different tasks can be different, and the following optimization problem is involved:

$$\begin{aligned} \min_{\{\mathbf{u}_\ell, \mathbf{v}_\ell\}, \Theta} \quad & \sum_{\ell=1}^m \left(\frac{1}{n_\ell} \sum_{i=1}^{n_\ell} L(\mathbf{u}_\ell^T x_i^\ell, y_i^\ell) + \alpha \|\mathbf{u}_\ell - \Theta^T \mathbf{v}_\ell\|^2 \right) \\ \text{s. t.} \quad & \Theta \Theta^T = I, \end{aligned} \quad (28)$$

where x_i^ℓ is the i th instance in the ℓ th task and n_ℓ is the number of instances in the ℓ th task. It has been shown [Ando and Zhang 2005] that the resulting optimization problem is non-convex even for convex loss functions. Hence, an iterative procedure called the *alternating structure optimization* (ASO) algorithm is proposed to compute a locally optimal solution. A similar idea of sharing parts of the model parameters among multiple tasks has been explored in the Bayesian framework [Bakker and Heskes 2003].

4.1.3 Multi-Class Learning. Formulation for extracting shared structures in multi-class classification has been proposed recently [Amit et al. 2007]. In this formulation, a low-rank transformation is computed to uncover the shared structures in multi-class classification. The final prediction is solely based on the low-dimensional representations in the dimensionality-reduced space. Moreover, the low-rank constraint is nonconvex, and it is first relaxed to the convex trace norm constraint. The relaxed problem can be formulated as a semidefinite program that is expensive to solve. Hence, the gradient-based optimization technique is employed to solve the relaxed problem.

4.2 Connections with Existing Formulations

The formulation proposed in Section 2 includes several existing algorithms as special cases. In particular, by setting the regularization parameters α and β in Equation (5) to different values, we obtain several well-known algorithms.

- $\alpha = 0$: When the regularization parameter $\alpha = 0$, it can be seen from Equation (5) that this formulation is equivalent to the classical ridge regression [Hoerl and Kennard 1970]. In ridge regression, different labels are decoupled, and the solution to each label can be obtained independently by solving a system of linear equations. In this case, no shared information is exploited among different labels.

- $\beta = 0$: When the regularization parameter $\beta = 0$, only the task-specific parameters $\{w_\ell\}_{\ell=1}^m$ are regularized. Thus, the proposed formulation reduces to the one in Ando and Zhang [2005] in the special case where the input data are the same for all tasks.
- $\alpha = +\infty$: It can be seen from Equation (20) that when α tends to infinity, the following holds:

$$\left(\frac{1}{n}X^T X + (\alpha + \beta)I\right)^{-1} \rightarrow \epsilon I,$$

for some small positive ϵ . Hence, the eigenvectors of $S_1^{-1}S_2$ approach the eigenvectors of the matrix

$$\left(\frac{1}{n}X^T X + \beta I\right)^{-1} X^T Y Y^T X. \quad (29)$$

This formulation is the same as the problem solved by orthonormalized PLS [Arenas-García et al. 2007]. When the matrix $Y Y^T$ in Equation (29) is replaced by $Y(Y^T Y)^{-1}Y^T$, this problem reduces to CCA. In the special case of multi-class problems, where each data point belongs to one class only, we define the class indicator matrix Y as follows [Ye 2007]: $y_{ij} = \sqrt{n/n_j} - \sqrt{n_j/n}$ if $y_i = j$, and $-\sqrt{n_j/n}$ otherwise, where n_j is the sample size of the j -th class. It is easy to verify that $\frac{1}{n}X^T X$ and $X^T Y Y^T X$ correspond to the total scatter and interclass scatter matrices used in LDA. Thus, the optimal Θ coincides with the optimal transformation computed by LDA.

- $\beta = +\infty$: When β tends to infinity, the eigenvectors of $S_1^{-1}S_2$ are given by the eigenvectors of the matrix $X^T Y Y^T X$, which is the interclass scatter matrix used in LDA. In this case, the proposed formulation is closely related to the orthogonal centroid method (OCM) [Park et al. 2003] in which the optimal transformation is given by the eigenvectors of the inter-class scatter matrix corresponding to the largest eigenvalues.

5. A FEATURE SPACE FORMULATION

In this section, we show that the proposed formulation can be extended to the kernel-induced feature space. Let $\Phi(X) = [\Phi(x_1), \dots, \Phi(x_n)]^T$ be the data matrix in the feature space induced by the feature mapping Φ . It follows from the Representer Theorem [Schölkopf and Smola 2002] that

$$U = \Phi(X)^{TA}, \quad (30)$$

$$\Theta = B^T \Phi(X), \quad (31)$$

for some matrices $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{n \times r}$. The feature space formulation of the proposed framework is summarized in the following theorem:

THEOREM 5.1. *Let A , B , and Y be defined as before, and let $K = \Phi(X)\Phi(X)^T$ be the kernel matrix. Then the optimal B^* in Equation (31) can be obtained by solving the following problem:*

$$\begin{aligned} \max_B \quad & tr((B^T \hat{S}_1 B)^{-1} B^T \hat{S}_2 B) \\ \text{s. t.} \quad & B^T K B = I, \end{aligned} \quad (32)$$

where N , \hat{S}_1 , and \hat{S}_2 are defined as

$$N = \frac{1}{n}KK + (\alpha + \beta)K, \quad (33)$$

$$\hat{S}_1 = K - \alpha KN^{-1}K, \quad (34)$$

$$\hat{S}_2 = KN^{-1}KYY^TKN^{-1}K. \quad (35)$$

PROOF. We first consider the term $\|U - \Theta^T V\|_F^2$ in Equation (5), which can be expressed in the feature space as

$$\text{tr}(A^T KA + V^T B^T KBV - 2A^T KBV). \quad (36)$$

Taking the derivative of the expression in Equation (36) with respect to V , and setting it to zero, we obtain

$$V = B^T KA. \quad (37)$$

Substituting this expression for V into Equation (5), and expressing all terms in the feature space, we get the following optimization problem with respect to A and B :

$$\begin{aligned} \min_{A,B} \quad & \frac{1}{n} \|KA - Y\|_F^2 + \text{tr}(A^T((\alpha + \beta)K - \alpha KBB^T K)A) \\ \text{s. t.} \quad & B^T KB = I. \end{aligned} \quad (38)$$

Taking the derivative of the objective function in Equation (38) with respect to A , and setting it to zero, we obtain the following expression for A :

$$A = \frac{1}{n}(N - \alpha KBB^T K)^{-1}KY, \quad (39)$$

where N is defined in Equation (33). Substituting this expression for A into the objective function in Equation (38), we get the following optimization problem with respect to B :

$$\begin{aligned} \max_B \quad & \frac{1}{n^2} \text{tr}(Y^T K(N - \alpha KBB^T K)^{-1}KY) \\ \text{s. t.} \quad & B^T KB = I. \end{aligned} \quad (40)$$

It follows from the Sherman-Woodbury-Morrison formula in Equation (16) that

$$(N - \alpha KBB^T K)^{-1} = N^{-1} + \alpha N^{-1}KB(B^T(K - \alpha KN^{-1}K)B)^{-1}B^T KN^{-1}.$$

This theorem can be proved by noticing the definitions of \tilde{S}_1 and \tilde{S}_2 in Equations (34) and (35), respectively. \square

It follows from the Sherman-Woodbury-Morrison formula in Equation (16) that

$$N^{-1} = \frac{1}{\alpha + \beta}K^{-1} - \frac{1}{\alpha + \beta}(n(\alpha + \beta)I + K)^{-1}. \quad (41)$$

Hence we have

$$K - \alpha KN^{-1}K = \frac{\beta}{\alpha + \beta}K + \frac{\alpha}{\alpha + \beta}K(n(\alpha + \beta)I + K)^{-1}K.$$

It follows that

$$\begin{aligned}\hat{\mathcal{S}}_1^{-1}\hat{\mathcal{S}}_2 &= (K - \alpha KN^{-1}K)^{-1}KN^{-1}KYY^TKN^{-1}K \\ &= (N - \alpha K)^{-1}KYY^TKN^{-1}K \\ &= \left(\frac{1}{n}KK + \beta K\right)^{-1}KYY^TK\left(\frac{1}{n}K + (\alpha + \beta)I\right)^{-1}.\end{aligned}$$

Similar to the discussion in Section 4, the connections between this formulation and related algorithms in the kernel-induced feature space can also be derived.

6. EXPERIMENTAL EVALUATION

In this section, we evaluate the effectiveness of the proposed formulations in comparison with representative algorithms. The linear formulation is evaluated on eleven multi-topic web page categorization problems, and the kernel formulation is evaluated on a series of gene expression pattern image annotation tasks.

6.1 Experimental Setup

We use area under the receiver operating characteristic (ROC) curve, called AUC, and the F1 score as the performance measure. To measure the performance across multiple labels using the F1 score, we use both the macro F1 and the micro F1 scores [Lewis et al. 2004; Yang and Pedersen 1997]. The F1 score depends on the threshold values of the classification models. It was shown recently [Fan and Lin 2007] that tuning the threshold based on F1 score on the training data can significantly improve performance. Hence, we tune the threshold value of each model based on the training data. Indeed, results show that threshold tuning sometimes outperforms classifiers that are trained to optimize the F1 score directly.

We evaluate the performance of the proposed linear and kernel formulations by comparing their performance with that of other five relevant methods. Parameters of all the methods are tuned using 5-fold cross-validation based on the F1 score. The setup is summarized as follows:

- MLLS*. The proposed multi-label formulation based on the least squares loss. The regularization parameters α and β are tuned using 5-fold double cross-validation from the candidate set $[0, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1]$. The performance of the proposed formulation is not sensitive to the dimensionality of the shared subspace r as long as it is not too small. We fix it to $5 \times \lfloor (m - 1)/5 \rfloor$ in the experiments where m is the number of labels.
- CCA+Ridge*. CCA is applied first to reduce the data dimensionality before ridge regression is applied. The regularization parameters for CCA and ridge regression are tuned on the sets $\{10^i | i = -6, -5, \dots, 1\}$ and $\{10^i | i = -6, -5, \dots, 0\}$, respectively.
- CCA+SVM*. CCA is applied first to reduce the data dimensionality before linear SVM is applied. The regularization parameter for CCA is tuned on the set $\{10^i | i = -6, -5, \dots, 1\}$, and the C value for SVM is tuned on the set $\{10^i | i = -4, -3, \dots, 4, 5\}$.

- SVM*. Linear SVM is applied on each label using the one-against-rest scheme, and the C value for each SVM is tuned on the set $\{10^i | i = -5, -4, \dots, 4, 5\}$.
- ASO_{SVM}*. The alternating structural optimization (ASO) algorithm proposed in [Ando and Zhang 2005] with the hinge loss as described in Equation (28). The regularization parameter α is tuned on the set $\{10^i | i = -4, -3, \dots, 2, 3\}$. The tolerance parameter for testing convergence is set to 10^{-3} , and the maximum number of iterations for ASO is set to 100. This problem is solved using the MOSEK package [Andersen and Andersen 2000].
- SVM^{perf}*. The SVM for multivariate performance measures proposed in [Joachims 2005, 2006]. We set $C = 5000$, as it leads to good overall performance.

The SVM problems are solved using the LIBSVM [Chang and Lin 2001] software package. All the codes and the web page categorization datasets used for the experiments are available from the supplemental web site (<http://www.public.asu.edu/~sji03/subspace>).

6.2 Web Page Categorization

The multi-topic web page categorization datasets were described in Ueda and Saito [2002a, 2002b] and Kazawa et al. [2005], and they were compiled from 11 top-level categories in the “yahoo.com” domain. The Web pages collected from each top-level category form a data set. The top-level categories are further divided into a number of second-level subcategories, and those subcategories form the topics to be categorized in each data set. Note that the 11 multi-topic categorization problems are compiled and solved independently as in Ueda and Saito [2002a]. We preprocess the datasets by removing topics with less than 100 web pages, words occurring less than 5 times, and Web pages without topics. We use the TF-IDF encoding to represent Web pages, and all Web pages are normalized to unit length. The statistics of all datasets are summarized in Table II.

6.2.1 Performance Evaluation. We randomly sample 1000 datapoints from each dataset as training data (each label is guaranteed to appear in at least one datapoint), and the remaining datapoints are used as test data. For the alternating structure optimization (ASO) algorithm [Ando and Zhang 2005] which is computationally expensive, we repeat this random sampling 10 times to generate 10 random training/test partitions. For all other methods, this random partitioning is repeated 30 times. Tables III and IV report the averaged performance of the six methods in terms of AUC, macro F1, and micro F1. We can observe that the proposed formulation achieves the highest AUC on seven datasets, while *ASO_{SVM}* achieves the highest AUC on the other four datasets. In terms of the macro F1 score, the proposed formulation achieves the highest performance on ten datasets while *ASO_{SVM}* achieves the highest performance on the remaining one. In terms of the micro F1 score, the proposed formulation outperforms other methods on all datasets. In general, methods that can capture the correlation among different labels, such as *ML_{LS}* and *CCA+SVM*, tend to yield higher performance than those that reduce the multi-label problem to

Table II.
 Statistics of the Yahoo! datasets. m , d , and N denote the number of labels, the data dimensionality, and the total number of instance, respectively, in the dataset after preprocessing. “MaxNPI”/“MinNPI” denotes the maximum/minimum number of positive instances for each topic (label)

Dataset	m	d	N	MaxNPI	MinNPI
Arts	19	17973	7441	1838	104
Business	17	16621	9968	8648	110
Computers	23	25259	12371	6559	108
Education	14	20782	11817	3738	127
Entertainment	14	27435	12691	3687	221
Health	14	18430	9109	4703	114
Recreation	18	25095	12797	2534	169
Reference	15	26397	7929	3782	156
Science	22	24002	6345	1548	102
Social	21	32492	11914	5148	104
Society	21	29189	14507	7193	113

Table III.
 Summary of performance for the six compared methods on the first six Yahoo! Datasets in terms of AUC (top section), macro F1 (middle section), and micro F1 (bottom section). All parameters of the six methods are used by cross-validation, and the averaged performance over 10 random sampling of training instances for ASO_{SVM} and 30 random sampling for all other methods is reported. The highest performance is highlighted for each dataSet

Dataset	Arts	Business	Computer	Education	Entertainment	Health
ML _{LS}	0.7611	0.8313	0.7912	0.7771	0.8282	0.8539
CCA+Ridge	0.7573	0.8253	0.7893	0.7568	0.8044	0.8557
CCA+SVM	0.7393	0.8003	0.7717	0.7420	0.7749	0.8450
SVM	0.7425	0.7973	0.7641	0.7548	0.8045	0.8439
ASO _{SVM}	0.7678	0.8261	0.7847	0.7446	0.8207	0.8621
SVM ^{perf}	0.7599	0.8185	0.7846	0.7710	0.8234	0.8554
ML _{LS}	0.3572	0.4026	0.3093	0.4044	0.4881	0.5971
CCA+Ridge	0.3176	0.3896	0.2940	0.3640	0.4249	0.5728
CCA+SVM	0.3217	0.3918	0.3006	0.3681	0.4319	0.5689
SVM	0.3374	0.3776	0.2961	0.3819	0.4653	0.5657
ASO _{SVM}	0.3568	0.3736	0.2873	0.3262	0.4344	0.5814
SVM ^{perf}	0.3361	0.3211	0.2579	0.3777	0.4656	0.4953
ML _{LS}	0.4700	0.7618	0.5529	0.4999	0.5844	0.6816
CCA+Ridge	0.4530	0.7596	0.5527	0.4498	0.5413	0.6775
CCA+SVM	0.4479	0.7434	0.5392	0.4296	0.5100	0.6646
SVM	0.4134	0.7150	0.4848	0.4560	0.5419	0.6349
ASO _{SVM}	0.4449	0.7384	0.4305	0.4322	0.5605	0.6754
SVM ^{perf}	0.4087	0.5892	0.3957	0.4378	0.5336	0.5997

a set of independent binary-class problems such as SVM. This shows that incorporation of the correlation information among different labels can improve performance, and the proposed formulation based on the least squares loss is effective in exploiting such information.

Table IV.

Summary of performance for the six compared methods on the last five Yahoo! datasets. See the caption of Table III for detailed explanations

Dataset	Recreation	Reference	Science	Social	Society
ML _{LS}	0.8126	0.8223	0.8287	0.8320	0.7276
CCA+Ridge	0.8113	0.8200	0.8099	0.8225	0.7216
CCA+SVM	0.7946	0.7604	0.7790	0.7627	0.7054
SVM	0.7922	0.8037	0.8025	0.7869	0.7101
ASO _{SVM}	0.8123	0.8340	0.8073	0.7942	0.7321
SVM ^{perf}	0.8109	0.8202	0.8236	0.8240	0.7229
ML _{LS}	0.4478	0.4066	0.4241	0.3637	0.3256
CCA+Ridge	0.4227	0.3483	0.3530	0.3303	0.2992
CCA+SVM	0.4289	0.3404	0.3660	0.3360	0.3073
SVM	0.4254	0.3848	0.3976	0.3444	0.3115
ASO _{SVM}	0.4136	0.4116	0.3397	0.3017	0.3023
SVM ^{perf}	0.4184	0.3680	0.3673	0.2978	0.3059
ML _{LS}	0.5287	0.6016	0.5210	0.6649	0.4853
CCA+Ridge	0.5197	0.5535	0.4719	0.6545	0.4813
CCA+SVM	0.5134	0.4435	0.4478	0.5688	0.4639
SVM	0.4753	0.5306	0.4565	0.5946	0.3971
ASO _{SVM}	0.4976	0.5580	0.4564	0.6492	0.4639
SVM ^{perf}	0.4620	0.4733	0.4155	0.4638	0.4034

Table V.

The p-values obtained by performing wilcoxon signed Rank test to assess the statistical significance of performance differences between ML_{LS} and Three other methods in terms of AUC (top section), macro F1 (middle section), and micro F1 (bottom section) on the first six

Yahoo! dataSets. A p-value of smaller than 0.05 is usually considered as indication of statistically significant difference

Dataset	Arts	Business	Computer	Education	Entertainment	Health
ML _{LS} v.s. CCA+Ridge	2.80e-1	5.31e-3	2.18e-2	9.31e-6	1.79e-5	4.16e-1
ML _{LS} v.s. CCA+SVM	7.51e-5	2.35e-6	7.15e-4	2.12e-6	1.73e-6	2.25e-3
ML _{LS} v.s. SVM	3.11e-5	1.92e-6	4.86e-5	2.35e-6	1.36e-5	4.44e-5
ML _{LS} v.s. CCA+Ridge	1.73e-6	2.22e-4	1.19e-3	1.92e-6	1.73e-6	3.06e-4
ML _{LS} v.s. CCA+SVM	1.73e-6	7.71e-4	2.18e-2	2.35e-6	1.73e-6	5.75e-6
ML _{LS} v.s. SVM	4.72e-6	1.63e-5	1.47e-2	5.21e-6	3.18e-6	2.60e-6
ML _{LS} v.s. CCA+Ridge	2.35e-6	1.71e-1	7.34e-1	1.73e-6	1.73e-6	9.36e-2
ML _{LS} v.s. CCA+SVM	4.72e-6	4.86e-5	1.14e-4	1.73e-6	1.92e-6	3.16e-3
ML _{LS} v.s. SVM	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6

To assess the statistical significance of performance differences between ML_{LS} and other compared methods, we perform Wilcoxon signed rank test based on the performance on 30 random trials, and the p-values are reported in Tables V and VI. A p-value of smaller than 0.05 is usually considered as indication of performance difference. We can observe that most of the performance differences are statistically significant. Note that ASO_{SVM} and SVM^{perf} are computationally expensive, and they require excessive amount of computational time to obtain the results on all 30 random trials, which is necessary for the purpose of statistical test. Hence, the test results for ASO_{SVM} and SVM^{perf} are omitted. However, we can observe that the performance of ASO_{SVM} and

Table VI.

The p-values obtained by performing wilcoxon signed rank test to assess the statistical significance of performance differences between ML_{LS} and three other methods in terms of AUC (top section), macro F1 (middle section), and micro F1 (bottom section) on the last five Yahoo! dataSets. A p-value of smaller than 0.05 is usually considered as indication of statistically significant difference

Dataset	Recreation	Reference	Science	Social	Society
ML_{LS} v.s. CCA+Ridge	5.98e-2	5.03e-1	2.22e-4	1.03e-3	3.28e-1
ML_{LS} v.s. CCA+SVM	3.58e-4	1.92e-6	2.35e-6	1.73e-6	3.72e-5
ML_{LS} v.s. SVM	5.79e-5	3.88e-4	3.88e-6	1.92e-6	6.31e-5
ML_{LS} v.s. CCA+Ridge	3.11e-5	1.73e-6	1.73e-6	2.87e-6	1.73e-6
ML_{LS} v.s. CCA+SVM	3.06e-4	1.73e-6	1.73e-6	2.35e-6	3.88e-6
ML_{LS} v.s. SVM	1.47e-4	4.44e-5	3.88e-6	3.06e-4	4.07e-5
ML_{LS} v.s. CCA+Ridge	7.15e-4	1.73e-6	1.73e-6	3.72e-5	4.65e-1
ML_{LS} v.s. CCA+SVM	9.62e-4	1.92e-6	1.73e-6	1.73e-6	3.50e-2
ML_{LS} v.s. SVM	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6

SVM^{perf} is usually lower than that of other methods, and hence we expect their performance differences with ML_{LS} are statistically significant.

6.2.2 Scalability Evaluation. We evaluate the scalability of the proposed multi-label formulation on all the Yahoo! datasets, and the results for 8 of the 11 datasets are presented in Figures 1 and 2. A similar trend can be observed from other datasets, and their results are omitted. In particular, we increase the number of training samples on the datasets gradually, and record the computation time of ML_{LS} , SVM, and ASO_{SVM} . The training time for a fixed parameter setting and the total time for parameter tuning using cross-validation are plotted in Figures 1 and 2. We can observe that SVM is the fastest and ASO_{SVM} is the slowest among the three compared algorithms. Moreover, the difference between ML_{LS} and SVM is small. The computational cost of the proposed formulation is dominated by the cost of SVD computation on the data matrix X , and it is independent of the number of labels. In contrast, the computational costs of SVM and ASO_{SVM} depend on the number of labels. Hence, the difference between SVM and ML_{LS} tends to be smaller on datasets with a larger number of labels. Note that in ML_{LS} , the two regularization parameters α and β are tuned using double cross-validation. However, the SVD on X needs to be computed only once irrespective of the size of the candidate sets for α and β . This experiment also shows that the running time of ASO_{SVM} may fluctuate as the number of training instances increases. This may be due to the fact that the convergence rate of the ASO_{SVM} algorithm depends on the initialization.

6.2.3 Sensitivity Analysis. We conduct experiments to evaluate the sensitivity of the proposed formulation to the values of the regularization parameters α and β . We randomly sample 1000 datapoints from each of the three datasets: Arts, Recreation, and Science, and the averaged macro F1 scores over 5-fold cross-validation for different values of α and β are depicted in Figure 3. We can observe that the highest performance on all three datasets is achieved at

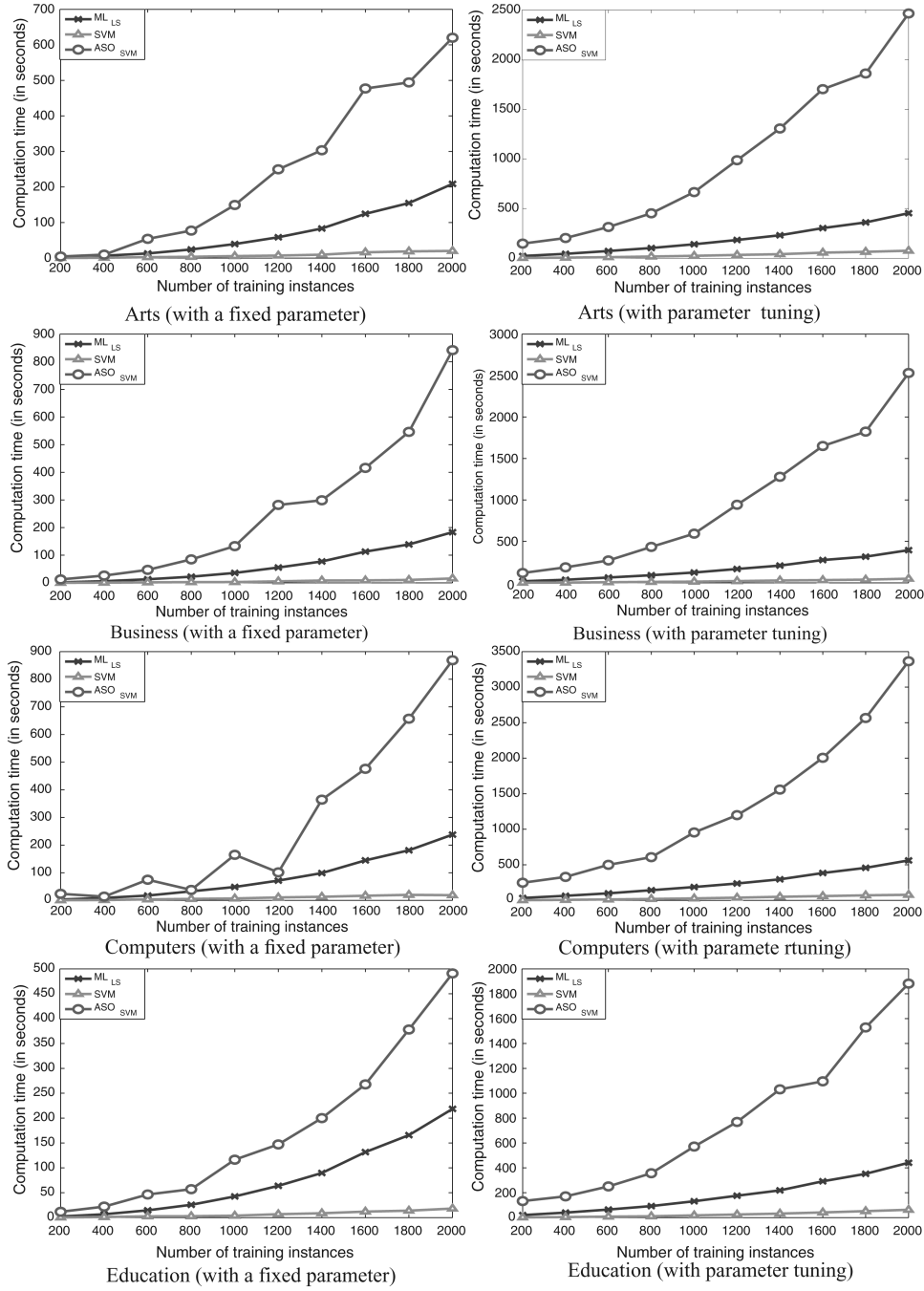


Fig. 1. Comparison of computation time for ML_{LS}, SVM, and ASO_{SVM} on four Yahoo! datasets. The computation time for a fixed parameter setting and that for parameter tuning using cross-validation are both depicted for each dataset. See the text for more details.

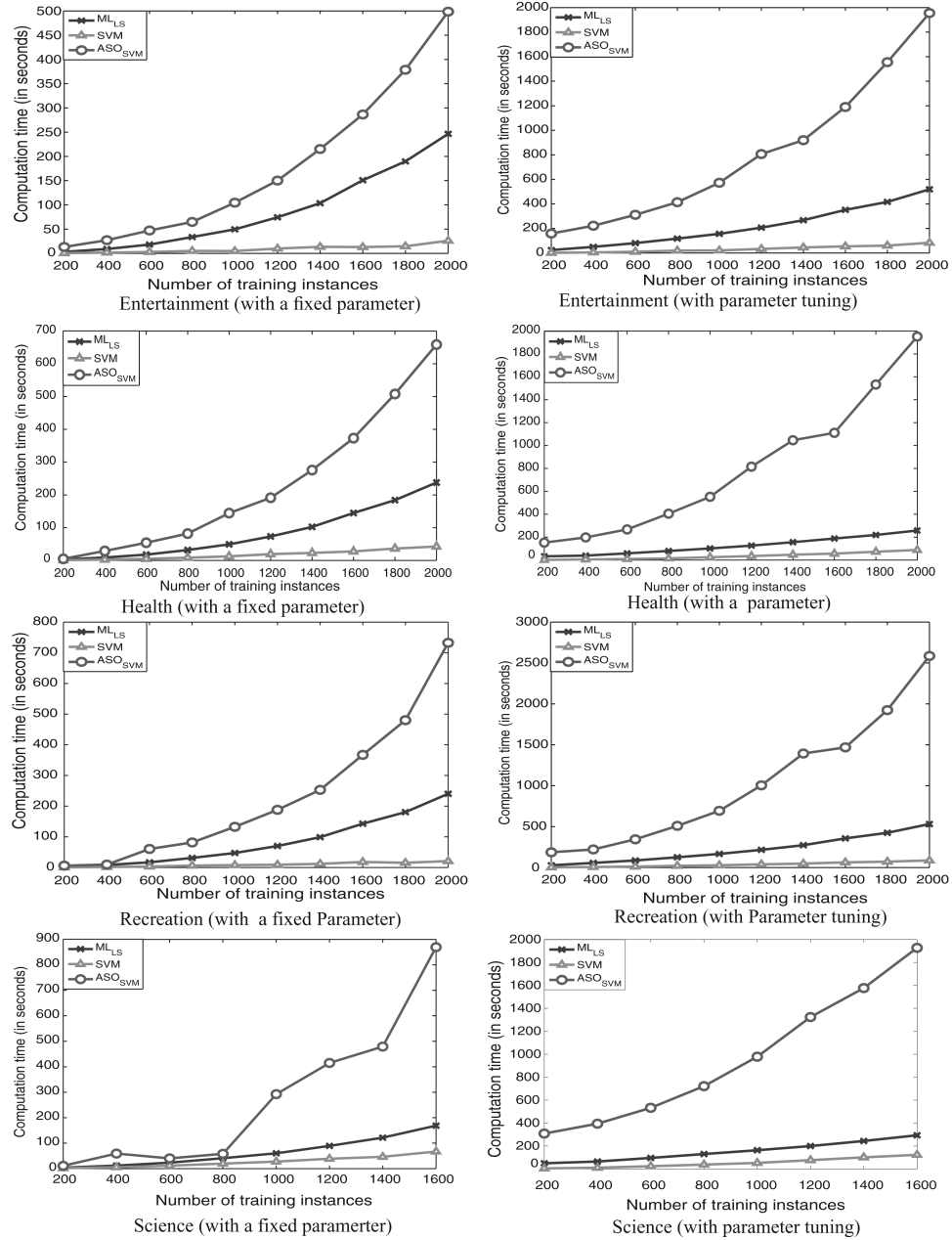


Fig. 2. Comparison of computation time for ML_{LS} , SVM, and ASO_{SVM} on four Yahoo! datasets. The computation time for a fixed parameter setting and that for parameter tuning using cross-validation are both depicted for each dataset. See the text for more details.

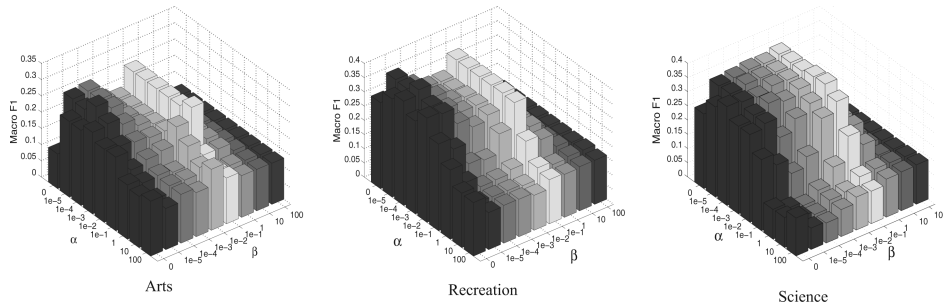


Fig. 3. The change of macro F1 scores as the regularization parameters α and β vary in the range $[0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100]$ for the Arts (left panel), Recreation (middle panel), and Science (right panel) datasets.

some intermediate values of α and β . Moreover, this experiment shows that the performance of the proposed multi-label formulation is sensitive to the values of the regularization parameters. Note that the parameter tuning time of the proposed formulation does not depend on the size of the candidate sets directly, since the computational cost is dominated by that of the SVD of X which needs to be performed only once. Hence, a large candidate set for α and β can be employed in practice.

6.3 Gene Expression Pattern Image Annotation

The gene expression pattern images of *Drosophila* document the spatial and temporal changes of gene expression during *Drosophila* embryogenesis [Tomancak et al. 2002; Tomancak et al. 2007] (Figure 4). Comparative analysis of such images can potentially reveal new genetic interactions, and yield insights into the complex regulatory networks governing embryonic development [Kumar et al. 2002]. To facilitate pattern comparison and searching, groups of images are annotated with a variable number of anatomical and developmental ontology terms using a controlled vocabulary in the Berkeley *Drosophila* Genome Project (BDGP) high-throughput study [Tomancak et al. 2002, 2007]. Since the number of available images produced by high-throughput technologies is rapidly increasing, it is imperative to design computational methods to automate this task [Ji et al. 2008, 2009a, 2009b; Li et al. 2009].

6.3.1 Kernel Matrix Construction. In Ji et al. [2008], a novel computational framework based on kernel methods is proposed to annotate gene expression pattern images automatically. In this framework, invariant features are first extracted from regular patches on each image in a group, resulting in a set of feature vectors for each image group [Mikolajczyk and Schmid 2005]. Note that although the number of local features extracted from each image is the same, the number of features extracted from different image groups may be different, since the number of images in each group may be different. The authors then propose to apply the vocabulary-guided pyramid match algorithm [Grauman and Darrell 2007, 2006] to construct kernels between groups of images based on the extracted features. A total of nine local descriptors are applied on regular

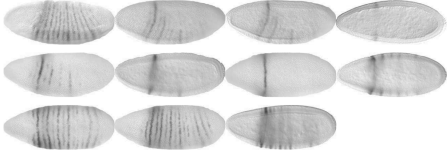
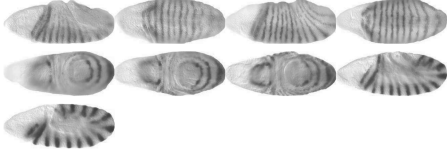
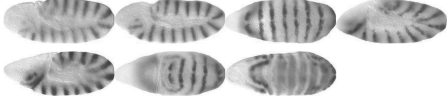
Stage range		Annotation terms
4-6		dorsal ectoderm anlage in situ nascendi mesectoderm anlage in situ nascendi segmentally repeated trunk mesoderm anlage in situ nascendi ventral ectoderm anlage in situ nascendi
7-8		dorsal ectoderm primordium hindgut anlage mesectoderm primordium procephalic ectoderm anlage trunk mesoderm primordium P2 ventral ectoderm primordium P2
9-10		inclusive hindgut primordium mesectoderm primordium procephalic ectoderm primordium trunk mesoderm primordium ventral ectoderm primordium

Fig. 4. Sample image groups and the associated terms in the FlyExpress database (<http://www.flyexpress.net>) for the segmentation gene *engrailed* in stage ranges 4–6, 7–8, and 9–10.

grids of 16 and 32 pixels in [Ji et al. 2008], resulting in 18 kernel matrices. In addition, five kernels are also constructed using global features computed from raw pixel values and Gabor filters. Since it has been shown [Ji et al. 2008] that kernels constructed from global descriptors yield lower performance than those constructed from local descriptors, we do not use these five kernels in this paper.

Motivated by the observation that integration of multiple feature types usually yields improved performance [Zhang et al. 2007], a multiple kernel learning formulation is proposed in [Ji et al. 2008] to combine the multiple candidate kernel matrices. In this experiment, we do not focus on multiple kernel learning. Instead, we combine the 18 kernels constructed from local descriptors in [Ji et al. 2008] with a uniform weight, as this strategy generates reasonably good kernels in practice [Tang et al. 2009]. The integrated kernel matrix is used to evaluate the feature space formulation proposed in Section 5.

6.3.2 Performance Evaluation. We use a collection of gene expression pattern images obtained from the FlyExpress database (<http://www.flyexpress.net/>) in the experiments. In particular, we select a number of terms from the FlyExpress database and extract a certain number of image groups annotated with at least one of the selected terms in the experiments. The number of terms used are 20, 30, 40, 50, and 60, and the number of image groups used is 1000 in the experiments. The extracted datasets are randomly partitioned into training and test sets using different ratios (3:7, 4:6, 5:5, and 6:4) for each label. The combination of five different numbers of terms and four different ratios results in a total of twenty sets of data. Similar to the setup on the Web page categorization task, the random partitioning of training and test sets is repeated 10 times for ASO_{SVM} and 30 times for all other methods, and the av-

Table VII.

Summary of performance for the six compared methods on the gene expression pattern image datasets in terms of AUC (top section), macro F1 (middle section), and micro F1 (bottom section). All parameters of the six methods are tuned by cross-validation, and the averaged performance over 10 random sampling of training instances for ASO_{SVM} and 30 random sampling for all other methods is reported. ‘ratio’ denotes the proportion of data used for training for each label. The highest performance is highlighted for each dataset

# of Terms	20		30		40		50		60	
Ratio	0.3	0.4	0.3	0.4	0.3	0.4	0.3	0.4	0.3	0.4
ML _{LS}	0.7778	0.7826	0.8307	0.8281	0.8312	0.7908	0.8058	0.8383	0.8395	0.8425
CCA+Ridge	0.7793	0.7943	0.8245	0.8249	0.8222	0.7925	0.8035	0.8376	0.8142	0.8311
CCA+SVM	0.7639	0.7581	0.7973	0.7909	0.7946	0.7764	0.7765	0.8110	0.8026	0.8030
SVM	0.7611	0.7741	0.8086	0.8005	0.8058	0.7700	0.7814	0.8139	0.8127	0.8149
ASO _{SVM}	0.7591	0.7706	0.8034	0.8064	0.7992	0.7629	0.7746	0.8127	0.8140	0.8093
ML _{LS}	0.4577	0.4005	0.4367	0.4057	0.3729	0.4690	0.4201	0.4297	0.4223	0.4034
CCA+Ridge	0.4578	0.4159	0.4344	0.3964	0.3807	0.4697	0.4244	0.4351	0.4070	0.3957
CCA+SVM	0.4481	0.3943	0.4244	0.3887	0.3720	0.4596	0.4089	0.4406	0.4045	0.3879
SVM	0.4444	0.4090	0.4084	0.3838	0.3653	0.4547	0.4138	0.4206	0.3998	0.3811
ASO _{SVM}	0.4432	0.4068	0.4003	0.3683	0.3507	0.4554	0.4135	0.4139	0.3917	0.3692
ML _{LS}	0.4923	0.4295	0.4626	0.4422	0.4133	0.5022	0.4515	0.4474	0.4536	0.4406
CCA+Ridge	0.4925	0.4500	0.4590	0.4318	0.4236	0.5026	0.4552	0.4548	0.4364	0.4321
CCA+SVM	0.4829	0.4337	0.4463	0.4165	0.4021	0.4937	0.4428	0.4573	0.4274	0.4134
SVM	0.4679	0.4245	0.4370	0.3943	0.3881	0.4700	0.4367	0.4293	0.4131	0.3975
ASO _{SVM}	0.4594	0.4203	0.4241	0.4088	0.3944	0.4644	0.4236	0.4314	0.4224	0.4022

eraged performance in terms of AUC, macro F1, and micro F1 is summarized in Tables VII and VIII.

We can observe from Tables VII and VIII that the proposed formulation achieves the highest AUC on seventeen out of the twenty sets of data. On the other three datasets, CCA+Ridge achieves the highest AUC. In terms of the macro F1 score, the proposed formulation and CCA+Ridge achieve the highest performance on twelve and eight datasets, respectively. In terms of the micro F1 score, the proposed formulation outperforms other five compared methods on ten datasets, while CCA+Ridge and CCA+SVM outperform other methods on nine and one datasets, respectively. Similar to the results observed on the web page categorization tasks, methods that incorporate the correlation among different labels consistently outperform those that reduce the multi-label problem into multiple independent binary-class problems. This again shows that the performance for multi-label problems can be improved by exploiting the correlation information among different labels. We can also observe from Tables VII and VIII that the performance difference between the proposed formulation and that of SVM tends to be larger when the number of labels increases. This may be due to the fact that when the number of labels is small, the correlation among different labels is weak and reducing the problem into independent binary-class problems may not compromise the performance significantly. However, when the number of labels is large, ignorance of the correlation information compromises the performance significantly. To assess the statistical significance of performance differences between ML_{LS} and other compared methods, we perform Wilcoxon signed rank test based on the performance on 30 random trials, and the p-values are reported in Tables IX and X. A p-value of smaller

Table VIII.

Summary of performance for the six compared methods on the gene expression pattern image datasets. The setup is the same as that reported in Table VII except that the training to test instance ratio are 5:5 and 6:4 in this table. See the caption of Table VII for detailed explanations

# of Terms	20		30		40		50		60	
Ratio	0.5	0.6	0.5	0.6	0.5	0.6	0.5	0.6	0.5	0.6
ML _{LS}	0.7994	0.8161	0.8492	0.8434	0.8500	0.8107	0.8211	0.8534	0.8490	0.8529
CCA+Ridge	0.7904	0.8119	0.8450	0.8369	0.8420	0.8101	0.8177	0.8464	0.8432	0.8448
CCA+SVM	0.7826	0.7822	0.8178	0.8064	0.8071	0.7961	0.7800	0.8216	0.8119	0.8106
SVM	0.7808	0.7897	0.8230	0.8183	0.8234	0.7899	0.7935	0.8284	0.8239	0.8246
ASO _{SVM}	0.7733	0.7881	0.8175	0.8146	0.8210	0.7760	0.8018	0.8347	0.8261	0.8245
ML _{LS}	0.4713	0.4380	0.4550	0.4256	0.4119	0.4818	0.4362	0.4657	0.4240	0.4122
CCA+Ridge	0.4637	0.4320	0.4412	0.4257	0.4108	0.4782	0.4330	0.4629	0.4279	0.4104
CCA+SVM	0.4612	0.4206	0.4468	0.4058	0.3966	0.4701	0.4106	0.4501	0.4076	0.3898
SVM	0.4522	0.4183	0.4282	0.4000	0.3901	0.4613	0.4216	0.4356	0.4070	0.3854
ASO _{SVM}	0.4579	0.4206	0.4181	0.3859	0.3860	0.4526	0.4289	0.4405	0.4058	0.3812
ML _{LS}	0.5006	0.4712	0.4721	0.4515	0.4430	0.5089	0.4682	0.4788	0.4464	0.4437
CCA+Ridge	0.4947	0.4586	0.4546	0.4541	0.4461	0.5066	0.4595	0.4765	0.4524	0.4440
CCA+SVM	0.4938	0.4533	0.4613	0.4257	0.4271	0.4992	0.4415	0.4616	0.4250	0.4135
SVM	0.4709	0.4340	0.4366	0.4125	0.3997	0.4801	0.4420	0.4378	0.4122	0.3957
ASO _{SVM}	0.4648	0.4299	0.4347	0.4064	0.4115	0.4598	0.4470	0.4492	0.4219	0.4082

Table IX.

The p-values obtained by performing wilcoxon signed rank test to assess the statistical significance of performance differences between ML_{LS} and three other methods in terms of AUC (top section), macro F1 (middle section), and micro F1 (bottom section) on the image annotation datasets. A p-value of smaller than 0.05 is usually considered as indication of statistically significant difference

# of Terms	20		30		40		50		60	
Ratio	0.3	0.4	0.3	0.4	0.3	0.4	0.3	0.4	0.3	0.4
ML _{LS} v.s. CCA+Ridge	1.15e-1	1.10e-1	1.73e-6	2.60e-6	1.92e-6	3.16e-2	1.73e-6	1.73e-6	1.73e-6	1.73e-6
ML _{LS} v.s. CCA+SVM	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6
ML _{LS} v.s. SVM	1.73e-6	1.73e-6	3.51e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6
ML _{LS} v.s. CCA+Ridge	8.61e-1	3.38e-1	1.92e-6	9.84e-3	6.26e-2	3.31e-4	1.02e-5	2.87e-6	5.70e-4	2.35e-6
ML _{LS} v.s. CCA+SVM	4.86e-5	3.58e-4	2.10e-3	2.12e-6	1.63e-5	4.44e-5	1.73e-6	1.73e-6	6.28e-1	1.73e-6
ML _{LS} v.s. SVM	5.75e-6	5.21e-6	3.72e-5	2.41e-3	1.73e-6	3.31e-4	1.73e-6	1.73e-6	1.03e-3	1.73e-6
ML _{LS} v.s. CCA+Ridge	9.42e-1	3.18e-1	1.73e-6	2.41e-3	1.75e-2	7.69e-6	3.18e-6	1.73e-6	5.21e-6	1.73e-6
ML _{LS} v.s. CCA+SVM	4.44e-5	1.05e-4	3.87e-2	5.75e-6	2.12e-6	1.02e-5	1.73e-6	1.73e-6	2.60e-6	1.73e-6
ML _{LS} v.s. SVM	1.73e-6	1.73e-6	1.47e-2	1.92e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6

than 0.05 is usually considered as indication of performance difference. We can observe that most of the performance differences are statistically significant.

6.4 Discussion

It follows from the discussion in Section 4 that ridge regression, CCA, and ASO_{SVM} with the same input data are all special cases of the proposed formulation. On both the web page categorization and the gene expression pattern image annotation tasks, the proposed formulations achieve the highest performance in most cases. An interesting observation is that the runner-up methods on these two tasks, which are ASO_{SVM} and CCA+Ridge, respectively, tend to be different. This may be attributable to the fact that the correlation information among different labels may be different for different

Table X.

The p-values obtained by performing wilcoxon signed rank test to assess the statistical significance of performance differences between ML_{LS} and three other methods in terms of AUC (top section), macro F1 (middle section), and micro F1 (bottom section) on the image annotation datasets. A p-value of smaller than 0.05 is usually considered as indication of statistically significant difference

# of Terms	20		30		40		50		60	
Ratio	0.5	0.6	0.5	0.6	0.5	0.6	0.5	0.6	0.5	0.6
ML_{LS} v.s. CCA+Ridge	1.73e-6	5.57e-1	1.73e-6	1.73e-6	2.35e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6
ML_{LS} v.s. CCA+SVM	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6
ML_{LS} v.s. SVM	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6
ML_{LS} v.s. CCA+Ridge	3.60e-3	1.25e-1	2.58e-3	4.94e-2	4.28e-6	1.20e-1	6.43e-1	2.70e-2	5.85e-1	1.65e-1
ML_{LS} v.s. CCA+SVM	8.18e-5	4.44e-5	1.73e-6	1.73e-6	1.73e-6	3.88e-4	1.73e-6	1.73e-6	2.35e-6	1.92e-6
ML_{LS} v.s. SVM	4.28e-6	1.92e-6	1.73e-6	1.92e-6	1.73e-6	1.73e-6	1.73e-6	2.35e-6	1.73e-6	1.73e-6
ML_{LS} v.s. CCA+Ridge	3.85e-3	1.52e-1	3.18e-6	2.37e-5	1.73e-6	1.77e-1	2.06e-2	3.88e-4	3.37e-3	7.65e-1
ML_{LS} v.s. CCA+SVM	2.83e-4	1.63e-5	1.92e-6	1.92e-6	6.98e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6
ML_{LS} v.s. SVM	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6	1.73e-6

tasks, and hence it should be exploited in different ways. One appealing feature of the proposed formulation is that it is a general framework that includes several well-known algorithms as special cases. By adjusting the two regularization parameters, the proposed formulation can be adapted to capture the correlation information among labels in various tasks. This effectively avoids the needs to apply multiple algorithms, such as ASO_{SVM} and CCA+Ridge, to a given task and choose the one that results in the highest performance.

The experimental results in this article show that the proposed shared-subspace formulation outperforms existing methods in most cases. This may be due to the fact that the proposed formulation is a general framework that includes several traditional algorithms, such as LDA, CCA, and PLS, as special cases. Hence, if the two regularization parameters are tuned properly, the proposed formulation is expected to outperform traditional methods, since it reduces to these methods when the regularization parameters are set to particular values. On the other hand, there are a few cases in the experiments in which the proposed formulation yields low performance. Recall that the proposed formulation assumes that a common subspace is shared among all labels, which may be too restrictive in some cases, and hence leads to low performance. Similar phenomenon has been observed in the contexts of multi-task learning [Argyriou et al. 2008; Jacob et al. 2009] and multivariate regression [Kim et al. 2008]. A commonly used technique to overcome this problem is to cluster the tasks into multiple clusters and impose local constraints onto tasks in the same cluster. We will extend the proposed formulation to clustered shared-subspace formulation and investigate its performance in the future.

7. CONCLUSIONS AND FUTURE WORK

We present a framework for extracting shared subspace in multi-label classification in this paper. In this framework, a subspace is assumed to be shared among multiple labels, and a linear transformation is computed to discover this subspace. We show that when the least squares loss is used in classification,

the optimal solution to the proposed formulation can be computed by solving a generalized eigenvalue problem. For high-dimensional data, direct computation is computationally expensive, and we develop an efficient algorithm for this case. We show that the proposed formulation is a general framework that includes several well-known formulations as special cases. Moreover, the proposed framework can be extended to the kernel-induced feature space. Experimental results on multi-topic Web page categorization and gene expression pattern image annotation tasks show that the proposed formulations outperform competing methods in most cases.

Our results show that applying regularization on both parts of the predictor can potentially improve performance. We have attempted to compare the proposed formulation with an extension of the ASO algorithm in which both parts of the predictor are regularized. However, this extension of the ASO algorithm is computationally demanding when both regularization parameters are tuned using double cross-validation. We will explore ways to improve the efficiency of this algorithm in the future. The data matrices in many applications such as the one in web page categorization task are sparse. Hence, techniques for computing the SVD of sparse matrices as proposed in Larsen [2000] can be employed to expedite the computation. We plan to apply such techniques in our algorithm in the future. The kernel matrix used in gene expression pattern image annotation is integrated from multiple candidate kernel matrices with a uniform weight. When some of the candidate kernel matrices are noisy, this simple approach for kernel integration may not yield improved performance. We plan to cast the proposed feature space formulation into the multiple kernel learning framework so that the weights for combining kernels can be adapted automatically.

REFERENCES

- AMIT, Y., FINK, M., SREBRO, N., AND ULLMAN, S. 2007. Uncovering shared structures in multi-class classification. In *Proceedings of the 24th International Conference on Machine Learning*. 17–24.
- ANDERSEN, E. D. AND ANDERSEN, K. D. 2000. The MOSEK interior point optimizer for linear programming: An implementation of the homogeneous algorithm. In *High Performance Optimization*. Kluwer Academic Publishers, 197–232.
- ANDO, R. K. AND ZHANG, T. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Resear.* 6, 1817–1853.
- ARENAS-GARCÍA, J., PETERSEN, K. B., AND HANSEN, L. K. 2007. Sparse kernel orthonormalized PLS for feature extraction in large data sets. *Adv. Neural Inform. Proces. Syst.* 19, 33–40.
- ARGYRIOU, A., MAURER, A., AND PONTIL, M. 2008. An algorithm for transfer learning in a heterogeneous environment. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*. 71–85.
- BAKKER, B. AND HESKES, T. 2003. Task clustering and gating for Bayesian multitask learning. *J. Mach. Learn. Resear.* 4, 83–99.
- BARNARD, K., DUYGULU, P., FORSYTH, D., D. FREITAS, N., BLEI, D. M., AND JORDAN, M. I. 2003. Matching words and pictures. *J. Mach. Learn. Resear.* 3, 1107–1135.
- BARUTCUOGLU, Z., SCHAPIRE, R. E., AND TROYANSKAYA, O. G. 2006. Hierarchical multi-label prediction of gene function. *Bioinformatics* 22, 7, 830–836.
- CARNEIRO, G., CHAN, A. B., MORENO, P. J., AND VASCONCELOS, M.-N. 2007. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. Patt. Anal. Mach. Intel.* 29, 3, 394–410.

- CHANG, C.-C. AND LIN, C.-J. 2001. *LIBSVM: A Library for Support Vector Machines*. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- ELISSEFF, A. AND WESTON, J. 2002. A kernel method for multi-labelled classification. *Adv. Neural Inform. Proces. Syst.* 14, 681–687.
- FAN, R.-E. AND LIN, C.-J. 2007. A study on threshold selection for multi-label classification. Tech. rep., Department of Computer Science and Information Engineering, National Taiwan University.
- FUKUNAGA, K. 1990. *Introduction to Statistical Pattern Recognition* 2nd Ed. Academic Press Professional.
- FUNG, G. M. AND MANGASARIAN, O. L. 2005. Multicategory proximal support vector machine classifiers. *Mach. Learn.* 59, 1-2, 77–97.
- GHAMRAWI, N. AND MCCALLUM, A. 2005. Collective multi-label classification. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*. 195–200.
- GOLUB, G. H. AND VAN LOAN, C. F. 1996. *Matrix Computations* 3rd Ed. The Johns Hopkins University Press.
- GRAUMAN, K. AND DARRELL, T. 2006. Approximate correspondences in high dimensions. *Adv. Neural Inform. Proces. Syst.* 19, 505–512.
- GRAUMAN, K. AND DARRELL, T. 2007. The pyramid match kernel: Efficient learning with sets of features. *J. Mach. Learn. Res.* 8, 725–760.
- HOERL, A. AND KENNARD, R. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 3, 55–67.
- HOTELLING, H. 1936. Relations between two sets of variates. *Biometrika* 28, 3-4, 321–377.
- JACOB, L., BACH, F., AND VERT, J.-P. 2009. Clustered multi-task learning: A convex formulation. *Adv. Neural Inform. Proces. Syst.* 21, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds. 745–752.
- JI, S., LI, Y.-X., ZHOU, Z.-H., KUMAR, S., AND YE, J. 2009a. A bag-of-words approach for *Drosophila* gene expression pattern annotation. *Bioinformatics* 10, 1, 119.
- JI, S., SUN, L., JIN, R., KUMAR, S., AND YE, J. 2008. Automated annotation of *Drosophila* gene expression patterns using a controlled vocabulary. *Bioinformatics* 24, 17, 1881–1888.
- JI, S. AND YE, J. 2009. Linear dimensionality reduction for multi-label classification. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*. 1077–1082.
- JI, S., YUAN, L., LI, Y.-X., ZHOU, Z.-H., KUMAR, S., AND YE, J. 2009b. *Drosophila* gene expression pattern annotation using sparse features and term-term interactions. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 407–416.
- JIN, R. AND GHAHRAMANI, Z. 2002. Learning with multiple labels. *Adv. Neural Inform. Proces. Syst.* 15, 897–904.
- JOACHIMS, T. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*. 137–142.
- JOACHIMS, T. 2005. A support vector method for multivariate performance measures. In *Proceedings of the 22nd International Conference on Machine Learning*. 377–384.
- JOACHIMS, T. 2006. Training linear SVMs in linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 217–226.
- KANG, F., JIN, R., AND SUKTHANKAR, R. 2006. Correlated label propagation with application to multi-label learning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1719–1726.
- KAZAWA, H., IZUMITANI, T., TAIRA, H., AND MAEDA, E. 2005. Maximal margin labeling for multi-topic text categorization. *Adv. Neural Inform. Proces. Syst.* 17, 649–656.
- KIM, S., SOHN, K.-A., AND XING, E. P. 2008. A multivariate regression approach to association analysis of quantitative trait network. Tech. rep. CMU-ML-08-113, Carnegie Mellon University.
- KUMAR, S., JAYARAMAN, K., PANCHANATHAN, S., GURUNATHAN, R., MARTI-SUBIRANA, A., AND NEWFELD, S. J. 2002. BEST: A novel computational approach for comparing gene expression patterns from early stages of *Drosophila melanogaster* development. *Genetics* 169, 2037–2047.
- LARSEN, R. M. 2000. Computing the SVD for large and sparse matrices. <http://soi.stanford.edu/~rmunk/PROPAC>.
- LEWIS, D. D., YANG, Y., ROSE, T. G., AND LI, F. 2004. RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.* 5, 361–397.

- LI, J. AND WANG, J. Z. 2008. Real-time computerized annotation of pictures. *IEEE Trans. Patt. Anal. Mach. Intel.* 3, 6, 985–1002.
- LI, Y.-X., JI, S., KUMAR, S., YE, J., AND ZHOU, Z.-H. 2009. *Drosophila* gene expression pattern annotation through multi-instance multi-label learning. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*. 1445–1450.
- MCCALLUM, A. 1999. Multi-label text classification with a mixture model trained by EM. In *Proceedings of the AAAI Workshop on Text Learning*.
- MIKOLAJCZYK, K. AND SCHMID, C. 2005. A performance evaluation of local descriptors. *IEEE Trans. Patt. Anal. Mach. Intel.* 27, 10, 1615–1630.
- MONAY, F. AND GATICA-PEREZ, D. 2007. Modeling semantic aspects for cross-media image indexing. *IEEE Trans. Patt. Anal. Mach. Intel.* 29, 10.
- PARK, H., JEON, M., AND ROSEN, J. B. 2003. Lower dimensional representation of text data based on centroids and least squares. *BIT* 43, 2, 1–22.
- RIFKIN, R. AND KLAUTAU, A. 2004. In defense of one-vs-all classification. *J. Mach. Learn. Res.* 5, 101–141.
- ROTH, V. AND FISCHER, B. 2007. Improved functional prediction of proteins by learning kernel combinations in multilabel settings. *Bioinformatics* 8, S12.
- SCHÖLKOPF, S. AND SMOLA, A. 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press.
- SUN, L., JI, S., AND YE, J. 2008a. Hypergraph spectral learning for multi-label classification. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- SUN, L., JI, S., AND YE, J. 2008b. A least squares formulation for canonical correlation analysis. In *Proceedings of the 25th International Conference on Machine Learning*. 1024–1031.
- SUN, L., JI, S., AND YE, J. 2009. On the equivalence between canonical correlation analysis and orthonormalized partial least squares. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*. 1230–1235.
- TANG, L., CHEN, J., AND YE, J. 2009. On multiple kernel learning with multiple labels. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*.
- TANG, L., RAJAN, S., AND NARAYANAN, V. K. 2009. Large scale multi-label classification via MetaLabeler. In *Proceedings of the 18th International World Wide Web Conference*.
- TOMANCAK, P., BEATON, A., WEISZMANN, R., KWAN, E., SHU, S. Q., LEWIS, S. E., RICHARDS, S., ASHBURNER, M., HARTENSTEIN, V., CELNIKER, S. E., AND RUBIN, G. 2002. Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* 3, 12.
- TOMANCAK, P., BERMAN, B., BEATON, A., WEISZMANN, R., KWAN, E., HARTENSTEIN, V., CELNIKER, S., AND RUBIN, G. 2007. Global analysis of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.* 8, 7, R145.
- UEDA, N. AND SAITO, K. 2002a. Parametric mixture models for multi-labeled text. *Adv. Neural Inform. Proces. Syst.* 15, 721–728.
- UEDA, N. AND SAITO, K. 2002b. Single-shot detection of multiple categories of text using parametric mixture models. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 626–631.
- WOLD, H. 1966. Estimation of principal components and related models by iterative least squares. P. R. Krishnaiah, Ed., *Multivariate Analysis*. Academic Press, New York, 391–420.
- YAN, R., TESIC, J., AND SMITH, J. R. 2007. Model-shared subspace boosting for multi-label classification. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 834–843.
- YANG, Y. AND PEDERSEN, J. O. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning*. 412–420.
- YE, J. 2005. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *J. Mach. Learn. Resear.* 6, 483–502.
- YE, J. 2007. Least squares linear discriminant analysis. In *Proceedings of the 24th International Conference on Machine Learning*. 1087–1093.

- YU, K., YU, S., AND TRESP, V. 2005. Multi-label informed latent semantic indexing. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. 258–265.
- ZHANG, J., MARSZALEK, M., LAZEBNIK, S., AND SCHMID, C. 2007. Local features and kernels for classification of texture and object categories: A comprehensive study. *Int. J. Comput. Vision* 73, 2, 213–238.
- ZHANG, M.-L. AND ZHOU, Z.-H. 2006. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Trans. Knowl. Data Engin.* 18, 10, 1338–1351.
- ZHANG, M.-L. AND ZHOU, Z.-H. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Patt. Recog.* 40, 7, 2038–2048.
- ZHOU, Z.-H. AND ZHANG, M.-L. 2007. Multi-instance multi-label learning with application to scene classification. *Adv. Neural Inform. Process. Syst.* 19. 1609–1616.

Received October 2008; revised April 2009; accepted August 2009