Linear Dimensionality Reduction for Multi-label Classification

Shuiwang Ji
Arizona State University
shuiwang.ji@asu.edu

Jieping Ye
Arizona State University
jieping.ye@asu.edu

Abstract

Dimensionality reduction is an essential step in high-dimensional data analysis. Many dimensionality reduction algorithms have been applied successfully to multi-class and multi-label problems. They are commonly applied as a separate data preprocessing step before classification algorithms. In this paper, we study a joint learning framework in which we perform dimensionality reduction and multi-label classification simultaneously. We show that when the least squares loss is used in classification, this joint learning decouples into two separate components, i.e., dimensionality reduction followed by multi-label classification. This analysis partially justifies the current practice of a separate application of dimensionality reduction for classification problems. We extend our analysis using other loss functions, including the hinge loss and the squared hinge loss. We further extend the formulation to the more general case where the input data for different class labels may differ, overcoming the limitation of traditional dimensionality reduction algorithms. Experiments on benchmark data sets have been conducted to evaluate the proposed joint formulations.

1 Introduction

Dimensionality reduction extracts a small number of features by removing irrelevant, redundant, and noisy information. It is a crucial step for the analysis of high-dimensional data. Classical dimensionality reduction techniques include unsupervised algorithms such as principal component analysis (PCA) [Jolliffe, 2002] and supervised algorithms such as linear discriminant analysis (LDA) [Fukunaga, 1990], canonical correlation analysis (CCA) [Hotelling, 1936], and partial least squares (PLS) [Arenas-García *et al.*, 2007]. These algorithms are commonly applied as a separate data preprocessing step before classification algorithms, and they have been applied successfully to many real-world problems.

One limitation of these approaches lies in the weak connection between dimensionality reduction and classification algorithms. Indeed, dimensionality reduction algorithms such as CCA and PLS and classification algorithms such as support vector machines (SVM) optimize different criteria. It is unclear which dimensionality reduction algorithm can best improve a specific classification algorithm such as SVM. In addition, most traditional dimensionality reduction algorithms assume that a common set of samples are involved for all classes. However, in many applications, e.g., when the data is unbalanced, it is desirable to relax this restriction so that the input data associated with each class can be better balanced. This is especially useful when some of the class labels in the data are missing.

In this paper we analyze dimensionality reduction in the context of multi-label classification [McCallum, 1999; Ueda and Saito, 2003; Zhang and Zhou, 2008]. We study a joint learning framework in which we perform dimensionality reduction and multi-label classification simultaneously. We show that when the least squares loss is used in classification, this joint learning decouples into two separate components, i.e., a separate dimensionality reduction step followed by multi-label classification. This partially justifies the current practice of a separate application of dimensionality reduction for classification problems. When other loss functions, including the hinge loss and the squared hinge loss, are employed the resulting optimization problems are non-convex. We show that they can be relaxed into convex-concave formulations. We propose a simple alternating algorithm to solve the joint learning problem. Experiments show that the alternating algorithm often converges in a few steps.

One appealing feature of the proposed joint learning formulations is that they can be extended naturally to cases where the input data for different labels may differ, overcoming the limitation of traditional dimensionality reduction algorithms. We conduct experiments using a collection of multi-label data sets. Results show that the joint formulations are comparable to a separate dimensionality reduction and classification, while they significantly outperform classification without dimensionality reduction. We demonstrate the superiority of the joint formulation using a data set in which the input data for different labels differ, and thus traditional dimensionality reduction algorithms are not applicable.

2 Background

In binary-class classification, we are given a data set $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$ is the input, $y_i \in \{-1, 1\}$ is the output, and n is the number of data points. We consider a linear classifier $f : x \in \mathbb{R}^d \to f(x) = \mathbf{w}^T x + b$ that minimizes the following regularized cost function:

$$E(f) = \sum_{i=1}^{n} L\left(y_i, f(x_i)\right) + \mu\Omega(f), \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the weight vector, $b \in \mathbb{R}$ is the bias, L is a prescribed loss function, Ω is a regularization functional measuring the smoothness of f, and $\mu > 0$ is the regularization parameter. Different loss functions lead to different learning algorithms.

In multi-label classification with k labels, each x_i can be associated with multiple labels, that is, $y_i \in \mathcal{P}(\{1, \cdots, k\})$ where $\mathcal{P}(\cdot)$ denotes the power set. The model in Eq. (1) can be extended to the multi-label case by constructing one binary classifier for each label in which instances relevant to this label form the positive class, and the rest form the negative class. This is an extension of the one-against-rest scheme commonly applied for multi-class classifications [Rifkin and Klautau, 2004].

2.1 Least Squares Loss

If the least squares loss is applied for multi-label classification, we compute a set of k linear functions, $f_\ell : x \to f_\ell(x) = \mathbf{w}_\ell^T x + b_\ell$, $\ell = 1, \cdots, k$, that minimize the following objective function:

$$E_1(\{f_\ell\}) = \sum_{\ell=1}^{k} \left(\sum_{i=1}^{n} \left(f_\ell(x_i) - y_{i\ell}\right)^2 + \mu||\mathbf{w}_\ell||_2^2 \right), \tag{2}$$

where $Y = (y_{i\ell}) \in \mathbb{R}^{n\times k}$ is the class label indicator matrix defined as: $y_{i\ell} = 1$ if $\ell \in y_i$, and -1 otherwise.

2.2 Hinge Loss

If the hinge loss is applied for multi-label classification, we consider a set of k linear functions, $f_\ell : x \to f_\ell(x) = \mathbf{w}_\ell^T x + b_\ell$, $\ell = 1, \cdots, k$, that minimize the following objective function:

$$E_2(\{f_\ell\}) = \sum_{\ell=1}^{k} \left(\sum_{i=1}^{n} \left(1 - f_\ell(x_i) y_{i\ell}\right)_+ + \mu||\mathbf{w}_\ell||_2^2 \right), \tag{3}$$

where $(z)_+ = \max(0, z)$. The ℓ-th linear function can be computed by minimizing $\sum_{i=1}^{n} (1 - f_\ell(x_i)y_{i\ell})_+ + \mu||\mathbf{w}_\ell||_2^2$, whose dual problem is given by

$$\begin{aligned} \max_{\boldsymbol{\alpha}^\ell \in \mathbb{R}^n} \quad & \sum_{i=1}^{n} \alpha_i^\ell - \frac{1}{2}(\boldsymbol{\alpha}^\ell)^T D^\ell X X^T D^\ell \boldsymbol{\alpha}^\ell \\ \text{s. t.} \quad & \sum_{i=1}^{n} y_{i\ell}\alpha_i^\ell = 0, \qquad 0 \le \boldsymbol{\alpha}^\ell \le C, \end{aligned} \tag{4}$$

where $X = [x_i, \cdots, x_n]^T$ is the data matrix, $C = \frac{1}{2\mu}$, $\boldsymbol{\alpha}^\ell \in \mathbb{R}^n$ is the vector of Lagrange dual variables, D^ℓ is a diagonal matrix with $(D^\ell)_{ii} = y_{i\ell}$. This is a standard quadratic programming (QP) problem.

2.3 Dimensionality Reduction

When the input data lie in a high-dimensional space, dimensionality reduction is commonly applied as a separate data preprocessing step. Principal component analysis (PCA) [Jolliffe, 2002] is a well-known technique for unsupervised dimensionality reduction. PCA reduces the data dimensionality while keeping the variance of the data as much as possible. Linear discriminant analysis (LDA) [Fukunaga, 1990] is a supervised dimensionality reduction technique in which the projection is obtained by maximizing the ratio of inter-class distance to intra-class distance. Canonical correlation analysis (CCA) and partial least squares (PLS) are commonly-used dimensionality reduction techniques for multi-label problems. All of these algorithms are applied as a separate preprocessing step before classification algorithms. In the following, we study a joint learning framework in which we perform dimensionality reduction and multi-label classification simultaneously.

3 Joint Dimensionality Reduction and Multi-label Classification

We study a joint learning framework for simultaneous dimensionality reduction and multi-label classification. In this framework, we learn a set of k linear functions, $f_\ell : x \to f_\ell(x) = \mathbf{w}_\ell^T Q^T x + b_\ell$, $\ell = 1, \cdots, k$, that minimize the following objective function:

$$E_3(\{f_\ell\}, Q) = \sum_{\ell=1}^{k} \left(\sum_{i=1}^{n} L(y_{i\ell}, f_\ell(x_i)) + \mu||\mathbf{w}_\ell||_2^2 \right), \tag{5}$$

where $Q \in \mathbb{R}^{d\times r}$ is the projection matrix, r is the reduced dimensionality, and $\mathbf{w}_\ell \in \mathbb{R}^r$ is the weight vector. We show that when the least squares loss is used, the joint optimization of Q and W results in a closed-form solution. Moreover, the optimal transformation is closely related to classical dimensionality reduction techniques discussed in Section 1.

3.1 Joint Learning with the Least Squares Loss

We assume that both the input X and the output Y are centered. In this case, all bias terms $\{b_\ell\}$ are zero, and the optimization problem in Eq. (5) becomes

$$\min_{W,Q:Q^TQ=I} ||XQW - Y||_F^2 + \mu||W||_F^2, \tag{6}$$

where $||\cdot||_F$ denotes the Frobenius norm [Golub and Van Loan, 1996] and $W = [\mathbf{w}_1, \cdots, \mathbf{w}_k]$. The optimal solution to the above optimization problem is given by a closed-form, as summarized in the following theorem:

Theorem 3.1. *Let Y be the target matrix defined from the labels. Then the optimal W that solves the joint learning problem in Eq. (6) is given by*

$$W = \left(Q^T X^T X Q + \mu I\right)^{-1} Q^T X^T Y, \tag{7}$$

and the optimal Q can be computed by solving

$$\max_{Q} tr\left(((Q^T(X^TX + \mu I)Q)^{-1}Q^TX^TYY^TXQ\right). \tag{8}$$

Proof. Taking the derivative of the objective in Eq. (6) with respect to W and setting it to zero, we have

$$W = \left(Q^T X^T X Q + \mu I\right)^{-1} Q^T X^T Y. \tag{9}$$

Substituting W in Eq. (9) into the objective function in Eq. (6), we obtain Eq. (8). $\square$

Theorem 3.1 shows that the transformation Q and the weight matrix W can be computed in a closed-form when the least squares loss is used. The solutions depend on the class label indicator matrix Y. We show below that the joint learning formulation is connected with traditional dimensionality reduction algorithms when different choices of Y are applied:

PLS: For multi-label problems, the class label indicator matrix Y defined in Section 2.1 can be used. In this case, the optimal Q from the joint formulation in Theorem 3.1 coincides with the optimal transformation in orthonormalized partial least squares (OPLS) [Arenas-García *et al.*, 2007].

CCA: When the class indicator matrix is set to $Y(Y^TY)^{-\frac{1}{2}}$, the problem in Eq. (8) can be expressed as:

$$\max_{Q} \operatorname{tr}\left(((Q^T(X^TX+\mu I)Q)^{-1}Q^TX^TY(Y^TY)^{-1}Y^TXQ\right),$$

which is the regularized canonical correlation analysis (CCA) formulation [Sun *et al.*, 2008].

LDA: 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: $y_{ij} = \sqrt{n/n_j} - \sqrt{n_j/n}$ 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 X^TX and X^TYY^TX correspond to the total scatter and inter-class scatter matrices used in LDA [Fukunaga, 1990]. Thus, the optimal Q from the joint formulation in Theorem 3.1 coincides with the optimal transformation computed by LDA.

The analysis above shows that, in the least squares case, the joint learning of dimensionality reduction (the transformation Q) and multi-label classification (the weight matrix W) is decoupled into two separate steps. In particular, the joint learning of Q and W is equivalent to computing transformation Q first by some dimensionality reduction algorithms such as LDA, CCA, and OPLS, and then apply classification in the dimensionality-reduced space. Therefore, performance is not expected to be improved by optimizing the transformation and the weight matrix jointly. This result justifies the current practice of a separate application of dimensionality reduction for classification.

3.2 Joint Learning with the Hinge Loss

When the hinge loss is employed in the joint learning formulation in Eq. (5), we obtain the following optimization problem:

$$\begin{aligned}\min_{\{\mathbf{w}_\ell, \xi_i^\ell\}, Q} \quad & \sum_{\ell=1}^{k}\left(\frac{1}{2}||\mathbf{w}_\ell||^2 + C\sum_{i=1}^{n}\xi_i^\ell\right) \\ \text{s. t.} \quad & y_{i\ell}(\mathbf{w}_\ell^T Q^T x_i + b_\ell) \geq 1 - \xi_i^\ell,\ \xi_i^\ell \geq 0,\ \forall\, i, \ell, \\ & Q^TQ = I,\end{aligned} \tag{10}$$

where ξ_i^ℓ is the slack variable for x_i in the ℓ-th model. The dual form of the problem in Eq. (10) is given by:

$$\begin{aligned}\min_{Q}\max_{\{\boldsymbol{\alpha}^\ell\}} & \sum_{\ell=1}^{k}\left(\sum_{i=1}^{n}\alpha_i^\ell - \frac{1}{2}\left((\boldsymbol{\alpha}^\ell)^T D^\ell X Q Q^T X^T D^\ell \boldsymbol{\alpha}^\ell\right)\right) \\ \text{s. t.} \quad & \sum_{i=1}^{n} y_i^\ell \alpha_i^\ell = 0,\ \ 0 \leq \boldsymbol{\alpha}^\ell \leq C,\ \ \forall\, \ell, \\ & Q^TQ = I.\end{aligned} \tag{11}$$

Convex-concave Relaxation

The objective and the constraint $Q^TQ = I$ in Eq. (11) are non-convex with respect to Q. We show in the following that this problem can be relaxed to a convex-concave formulation. Specifically, we replace QQ^T with Z in the objective in Eq. (11) and add $QQ^T = Z$ to the constraint. It can be shown [Overton, 1993] that the set $\mathcal{Z} = \{Z|\text{tr}(Z) = r,\ 0 \preceq Z \preceq I\}$ is the convex hull of the non-convex set $\mathcal{Z}_0 = \{Z|Z = QQ^T, Q^TQ = I, Q \in \mathbb{R}^{d\times r}\}$, where $A \preceq B$ denotes that $B - A$ is positive semidefinite. Thus, the optimization problem in Eq. (11) can be relaxed to the following convex-concave problem:

$$\begin{aligned}\max_{\{\boldsymbol{\alpha}^\ell\}}\min_{Z} & \sum_{\ell=1}^{k}\left(\sum_{i=1}^{n}\alpha_i^\ell - \frac{1}{2}\left((\boldsymbol{\alpha}^\ell)^T D^\ell X Z X^T D^\ell \boldsymbol{\alpha}^\ell\right)\right) \\ \text{s. t.} \quad & \sum_{i=1}^{n} y_i^\ell \alpha_i^\ell = 0,\ 0 \leq \boldsymbol{\alpha}^\ell \leq C,\ \forall\ \ell, \\ & \text{tr}(Z) = r,\ 0 \preceq Z \preceq I.\end{aligned} \tag{12}$$

All the constraints in the formulation in Eq. (12) are convex. In addition, the objective is convex in Z and concave in $\{\boldsymbol{\alpha}^\ell\}_{\ell=1}^k$. Thus, this optimization problem is a convex-concave problem and the existence of a saddle point is guaranteed by the well-known von Neumann Lemma [Nemirovski, 1994]. Since the objective function is maximized in terms of $\{\boldsymbol{\alpha}^\ell\}_{\ell=1}^k$ and minimized in terms of Z at the saddle point, it is also the globally optimal solution to this problem [Nemirovski, 1994].

An Alternating Algorithm

We propose to solve the joint learning formulation in Eq. (11) iteratively. More specifically, when Q is fixed, solutions to $\{\boldsymbol{\alpha}^\ell\}_{\ell=1}^k$ are decoupled for different ℓ. Each $\boldsymbol{\alpha}^\ell$ can be obtained by solving a standard SVM problem with a modified kernel XQQ^TX^T. When $\{\boldsymbol{\alpha}^\ell\}_{\ell=1}^k$ is fixed, Q can be computed by solving the following problem:

$$\max_{Q:Q^TQ=I} \text{tr}(Q^TX^TSXQ), \tag{13}$$

where

$$S = \sum_{\ell=1}^{k}\left(D^\ell\boldsymbol{\alpha}^\ell(\boldsymbol{\alpha}^\ell)^TD^\ell\right). \tag{14}$$

It is known that this trace maximization problem has a closed-form solution. In particular, columns of the optimal Q^* consist of the left singular vectors of the matrix $[X^TD_1\boldsymbol{\alpha}^1, \cdots, X^TD_k\boldsymbol{\alpha}^k] \in \mathbb{R}^{d\times k}$. Experiments in Section 5 show that the proposed iterative procedure converges in a small number of steps.

It is interesting to note that the matrix S in Eq. (14) can be considered as a similarity matrix between data points. More specifically, the similarity between x_i and x_j is based on the vectors of Lagrangian variables $\{\boldsymbol{\alpha}^\ell\}$ computed from k SVMs, as well as their class label information in $\{D^\ell\}$. Intuitively, for x_i and x_j, if $y_{i\ell}\alpha_i^\ell$ is similar to $y_{j\ell}\alpha_j^\ell$ for all $\ell = 1, \cdots, k$, then these two points should have a high similarity score. Therefore, the computation of $\{\boldsymbol{\alpha}^\ell\}$ from k separate SVMs can be interpreted as an intermediate step of constructing a similarity matrix, which is subsequently used to compute the low-dimensional embedding.

Learning Orthonormal Features

In the above formulations, we require the transformation to be orthonormal, that is $Q^T Q = I$. We can also require the transformed features to be orthonormal by imposing the following constraint:

$$Q^T(X^T X + \mu I)Q = I, \tag{15}$$

where a regularization term is added to deal with the singularity problem of the covariance matrix. It can be shown that this constraint can also be relaxed to convex ones, resulting in a convex-concave problem. Similarly, the proposed iterative procedure can be adapted to solve this problem in which the iterative step for solving Q becomes

$$\max_Q \text{tr}\left(Q^T X^T S X Q\right),$$

subject to the constraint in Eq. (15). The optimal Q can be readily computed via solving a generalized eigenvalue problem.

Joint Learning with the Squared Hinge Loss

The squared hinge loss is also commonly used in SVM, which is defined as $L(y, f) = \max(0, 1 - yf)^2$. With this loss, the optimization problem in Eq. (5) becomes:

$$\begin{aligned} \min_{\{\mathbf{w}_\ell, \xi_i^\ell\}, Q} \quad & \sum_{\ell=1}^k \left(\frac{1}{2}||\mathbf{w}_\ell||^2 + C\sum_{i=1}^n \left(\xi_i^\ell\right)^2\right) \\ \text{s. t.} \quad & y_{i\ell}((\mathbf{w}_\ell)^T Q^T x_i + b_\ell) \geq 1 - \xi_i^\ell, \ \forall i, \ \ell, \\ & Q^T Q = I. \end{aligned} \tag{16}$$

The dual form of the problem in Eq. (16) is given by:

$$\begin{aligned} \min_Q \max_{\{\boldsymbol{\alpha}^\ell\}} \quad & \sum_{\ell=1}^k \left(\sum_{i=1}^n \alpha_i^\ell - \frac{1}{2}\left((\boldsymbol{\alpha}^\ell)^T \right.\right. \\ & \left.\left. \left(D^\ell X Q Q^T X^T D^\ell + \frac{1}{2C} I\right)\boldsymbol{\alpha}^\ell\right)\right) \\ \text{s. t.} \quad & \sum_{i=1}^n y_i^\ell \alpha_i^\ell = 0, \ \boldsymbol{\alpha}^\ell \geq 0, \ \forall \ \ell, \\ & Q^T Q = I. \end{aligned} \tag{17}$$

Similar techniques can be applied to relax the problem into a convex-concave formulation and derive an iterative algorithm to compute the solution. It can also be extended to learn orthonormal features as discussed above.

Related Work

Our joint learning formulation in Eq. (11) is closely related to the sparse learning algorithm proposed in [Wu *et al.*, 2006], which works on binary-class problems. The column vectors of Q are considered as pseudo support vectors in [Wu *et al.*, 2006], and no orthonormality condition is imposed on Q. In addition, [Wu *et al.*, 2006] focuses on constructing an approximate SVM by using a small set of support vectors, while we focus on dimensionality reduction embedded in SVM. Joint structure learning and classification for multi-task learning has been studied in [Ando and Zhang, 2005]. [Amit *et al.*, 2007] proposed joint feature extraction and multi-class SVM classification using the low-rank constraint. Due to the intractability of this constraint, it is relaxed to the trace norm constraint and the relaxed problem was solved by gradient-descent algorithms. The computations involved in the proposed formulation are much simpler than all these approaches, while our experiments below show that this simple iterative algorithm often achieves the globally optimal solution. [Argyriou *et al.*, 2007] proposed to learn a common sparse representation from multiple related tasks based on an iterative procedure, which is shown to converge to a global optimum. The analysis in [Argyriou *et al.*, 2007] may be used to prove the convergence property of a perturbed version of the proposed algorithm. In our formulation, the step for computing $\boldsymbol{\alpha}^\ell$ is reduced to solving a standard SVM with a modified kernel, and hence they are also related to the problem of kernel learning [Lanckriet *et al.*, 2004].

4 Dimensionality Reduction with Different Input Data

Our discussions above assume that the input data for all labels are the same, i.e., a common data matrix X for all labels. In many practical applications, especially when the data is unbalanced, it is desirable to relax this restriction so that the input data associated with each label can be better balanced. Traditional dimensionality reduction algorithms such as LDA, CCA, and PLS cannot be applied in such scenario. We show that the proposed joint formulations can be extended naturally to deal with such type of data.

Let X^ℓ be the data matrix of the ℓ-th label. We obtain the following optimization problem (in the dual form) under the hinge loss:

$$\begin{aligned} & \min_Q \max_{\{\boldsymbol{\alpha}^\ell\}} \sum_{\ell=1}^k (\sum_{i=1}^n \alpha_i^\ell - \frac{1}{2}((\boldsymbol{\alpha}^\ell)^T D^\ell X^\ell Q Q^T (X^\ell)^T D^\ell \boldsymbol{\alpha}^\ell)) \\ & \quad \text{s. t.} \quad \sum_{i=1}^n y_i^\ell \alpha_i^\ell = 0, \quad 0 \leq \boldsymbol{\alpha}^\ell \leq C, \quad \forall \ \ell, \ Q^T Q = I. \end{aligned}$$

Similar to the above discussions, the optimization problem in Eq. (18) can be solved iteratively. In particular, when Q is fixed, $\{\boldsymbol{\alpha}^\ell\}_{\ell=1}^k$ can be computed by solving k standard SVM problems with the kernel modified as $X^\ell Q Q^T (X^\ell)^T$. When $\{\boldsymbol{\alpha}^\ell\}_{\ell=1}^k$ is fixed, the following trace maximization problem is involved:

$$\max_{Q: Q^T Q = I} \text{tr}\left(Q^T \sum_{\ell=1}^k \left((X^\ell)^T D^\ell \boldsymbol{\alpha}^\ell (\boldsymbol{\alpha}^\ell)^T D^\ell X^\ell\right) Q\right). \tag{18}$$

Table 1: Mean ROC achieved by various formulations on the *art* (top) and *business* (bottom) data sets. The data sets are partitioned into training and test sets with different ratios, and the mean ROC values and standard deviations over ten random trials are reported in each case.

RATIO	MLSVM$^T_{L_1}$	MLSVM$^F_{L_1}$	MLSVM$^T_{L_2}$	MLSVM$^F_{L_2}$	CCA+SVM	SVM
20%	63.07±0.92	62.71±0.98	63.07±0.92	62.71±0.99	63.02±1.06	44.07±5.12
30%	64.15±0.60	63.55±0.98	64.15±0.60	63.51±0.96	63.72±0.96	49.61±3.47
40%	65.11±0.76	64.32±0.67	65.11±0.76	64.33±0.66	64.65±0.62	53.94±3.59
50%	65.74±0.67	65.05±1.11	65.74±0.67	65.04±1.13	65.17±1.00	56.92±3.93
60%	66.34±0.76	64.73±1.00	66.34±0.76	64.76±0.99	65.01±0.97	59.01±2.15
20%	68.74±3.56	70.89±1.99	68.74±3.56	70.89±2.00	71.06±2.02	38.39±6.81
30%	74.54±0.69	73.15±1.47	74.54±0.69	73.14±1.48	73.20±1.43	49.07±6.84
40%	75.33±0.91	74.08±1.36	75.33±0.91	74.09±1.37	74.17±1.31	59.82±4.88
50%	76.82±1.34	74.67±1.22	76.82±1.33	74.67±1.22	74.72±1.22	62.11±8.53
60%	77.69±1.47	76.07±1.38	77.69±1.47	76.05±1.37	76.15±1.42	68.59±5.64

It is known that columns of Q that solves the above problem consist of the left singular vectors of the matrix $[(X^1)^T D^1 \boldsymbol{\alpha}^1, \cdots, (X^k)^T D^k \boldsymbol{\alpha}^k] \in \mathbb{R}^{d \times k}$.

5 Experiments

In this section we evaluate the proposed formulations when the input data for different labels are the same or different.

5.1 Experiments on Multi-label Data Sets

The two multi-label data sets used are the *art* and *business*, which were originally used in [Ueda and Saito, 2003], and they consist of web pages from the art and business directories at Yahoo!. Each web page is assigned a variable number of labels indicating its categories. All instances are encoded with TF-IDF and are normalized to have unit length. These data sets are high-dimensional (23146 and 21924 dimensions), and we extract 20 labels and 1000 instances from each data set. We also conduct experiments on two other data sets *scene* and *yeast* and the detailed results are omitted due to space constraints, but the results are briefly summarized below. We report the receiver operating characteristic (ROC) values of the proposed four formulations in Table 1. The hinge loss and squared hinge loss multi-label SVM formulations with the orthonormal transformation and orthonormal features are denoted as MLSVM$^T_{L_1}$, MLSVM$^F_{L_1}$, MLSVM$^T_{L_2}$, and MLSVM$^F_{L_2}$, respectively. The performance of SVM in the original data space and in the dimensionality-reduced space by CCA (CCA+SVM) is also reported.

We observe from the results that the proposed formulations with the orthonormal transformation and orthonormal features achieve the highest performance on the two high-dimensional (*art* and *business*) and two low-dimensional (*scene* and *yeast*) data sets, respectively. The improvement over CCA+SVM is small on the four data sets. This implies that the joint learning of dimensionality reduction and classification is similar to applying them separately in some cases. The experiments also show that formulations based on dimensionality reduction generally outperform those in the original space, especially when the data dimensionality is high. This justifies the use of dimensionality reduction in multi-label classification.

To evaluate the convergence of the proposed iterative algorithm, we plot the objective values of the MLSVM$^T_{L_1}$ formulation on the *art* and *business* data sets after each update of Q and $\{\boldsymbol{\alpha}\}_{\ell=1}^k$ separately in Figure 1. We can see that the objective values of the maximization and minimization problems converge to the same point in a few steps on both data sets.

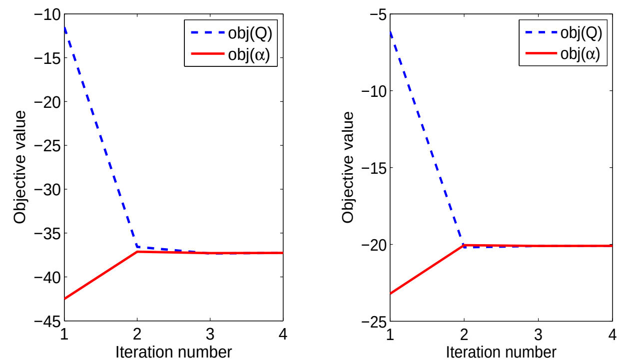

Figure 1: Convergence of MLSVM$^T_{L_1}$ on the *art* (left) and *business* (right) data sets. "obj(Q)" and "obj(α)" denote the objective values after updating Q and $\{\boldsymbol{\alpha}\}_{\ell=1}^k$, respectively, at each iteration.

5.2 Experiments on Data with Different Inputs

The *landmine* data [Xue *et al.*, 2007] consists of 29 subsets (tasks) that are collected from various landmine fields. Each object in a given task is represented by a 9-dimensional feature vector and a binary label indicating landmine or clutter. The inputs for different tasks are different. We apply MLSVM$^T_{L_1}$ on the *landmine* data to learn a common transformation for all of the tasks, and project them into a low-dimensional space using this transformation. This transformation can capture the common structures shared by all of the tasks and improve the detection performance. We also apply SVM on each of the task independently. The data for each task are partitioned into training and test sets with different proportions, and the averaged ROC values and standard deviations over 50 random partitions in each case are depicted in Figure 2. We can see that the proposed formulation can improve performance consistently by capturing the common predictive structures shared among multiple tasks.

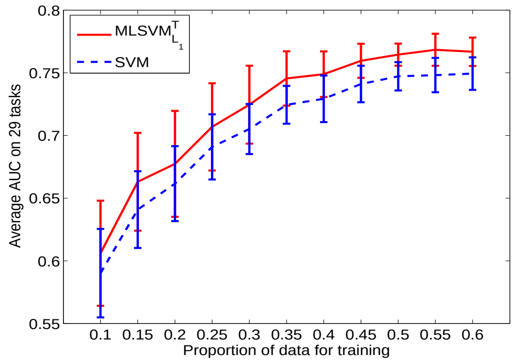

Figure 2: Average ROC for the landmine detection problem. Different proportions (indicated by x-axis) of the data are used for training, and the average ROC values over 50 random partitions are plotted.

To test the statistical significance of the differences between the performance of these two methods, we perform Wilcoxon signed rank test for the null hypothesis that the performance achieved by these two methods across 50 random trials is the same, and the maximum p-value obtained for different ratios of training/test splitting is 0.0097. This shows that the performance differences between these two methods are statistically significant. Note that traditional dimensionality reduction algorithms are not applicable for this problem.

6 Conclusion and Discussion

We study the role of dimensionality reduction in multi-label classification in this paper. We show that when the least squares loss is used in classification, the joint learning decouples into two separate components. When the hinge loss is used, the resulting optimization problems are non-convex, and we show that they can be relaxed into convex-concave formulations. We further extend the proposed formulations to the case where the input data for different labels may be different.

Experiments show that the proposed iterative algorithm converges in a small number of iterations. We plan to study this convergence property by using results developed in related fields [Argyriou *et al.*, 2007]. The relative performance of formulations with orthonormal transformation and orthonormal features is different for different data sets. We plan to analyze this in the future.

Acknowledgements

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

References

[Amit *et al.*, 2007] Y. Amit, M. Fink, N. Srebro, and S. Ullman. Uncovering shared structures in multiclass classification. In *ICML*, pages 17–24, 2007.

[Ando and Zhang, 2005] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

[Arenas-García *et al.*, 2007] J. Arenas-García, K. B. Petersen, and L. K. Hansen. Sparse kernel orthonormalized PLS for feature extraction in large data sets. In *NIPS*, pages 33–40. 2007.

[Argyriou *et al.*, 2007] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 2007.

[Fukunaga, 1990] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Press Professional, 2nd edition, 1990.

[Golub and Van Loan, 1996] G. H. Golub and C F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 3rd edition, 1996.

[Hotelling, 1936] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377, 1936.

[Jolliffe, 2002] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 2nd edition, 2002.

[Lanckriet *et al.*, 2004] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[McCallum, 1999] A. McCallum. Multi-label text classification with a mixture model trained by EM. In *AAAI Workshop on Text Learning*, 1999.

[Nemirovski, 1994] A. Nemirovski. Efficient methods in convex programming, 1994. Lecture Notes.

[Overton, 1993] M. L. Overton. Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Math. Programming*, 62(1):321–357, 1993.

[Rifkin and Klautau, 2004] Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004.

[Sun *et al.*, 2008] L. Sun, S. Ji, and J. Ye. A least squares formulation for canonical correlation analysis. In *ICML*, pages 1024–1031, 2008.

[Ueda and Saito, 2003] N. Ueda and K. Saito. Parametric mixture models for multi-labeled text. In *NIPS*, pages 721–728. 2003.

[Wu *et al.*, 2006] M. Wu, B. Schölkopf, and G. Bakir. A direct method for building sparse kernel learning algorithms. *Journal of Machine Learning Research*, 7:603–624, 2006.

[Xue *et al.*, 2007] Y. Xue, X. Liao, L. Carin, and B. Krishnapuram. Multi-task learning for classification with Dirichlet process priors. *Journal of Machine Learning Research*, 8:35–63, 2007.

[Zhang and Zhou, 2008] Y. Zhang and Z.-H. Zhou. Multi-label dimensionality reduction via dependency maximization. In *AAAI*, pages 1503–1505, 2008.