Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

Multi-Label Classification with Feature-Aware Non-Linear Label Space Transformation

Xin Li and **Yuhong Guo**
Department of Computer and Information Sciences
Temple University, Philadelphia, PA 19122, USA
{xinli, yuhong}@temple.edu

Abstract

Multi-label classification with many classes has recently drawn a lot of attention. Existing methods address this problem by performing linear label space transformation to reduce the dimension of label space, and then conducting independent regression for each reduced label dimension. These methods however do not capture nonlinear correlations of the multiple labels and may lead to significant information loss in the process of label space reduction. In this paper, we first propose to exploit kernel canonical correlation analysis (KCCA) to capture nonlinear label correlation information and perform nonlinear label space reduction. Then we develop a novel label space reduction method that explicitly combines linear and nonlinear label space transformations based on CCA and KCCA respectively to address multi-label classification with many classes. The proposed method is a feature-aware label transformation method that promotes the label predictability in the transformed label space from the input features. We conduct experiments on a number of multi-label classification datasets. The proposed approach demonstrates good performance, comparing to a number of state-of-the-art label dimension reduction methods.

1 Introduction

Multi-label classification is an important problem in many application domains, where each data instance can be assigned into multiple categories. For example, in image labeling [Zhou and Zhang, 2006] and video annotation [Qi *et al.*, 2007], an image can contain multiple objects and thus have multiple labels from a large number of object classes. In text categorization [Schapire and Singer, 2000], a given article or webpage can be assigned into multiple topics. A simple way of multi-label classification transforms the multi-label learning problem into a set of independent single label classification problems [Lewis *et al.*, 2004; Chen *et al.*, 2007]. This type of methods however have the obvious drawback of ignoring the critical correlation information between the multiple labels. A significant number of multi-label learning works developed in the literature have centered on exploiting the label interdependency information between the multiple labels [Elisseeff and Weston, 2001; Guo and Gu, 2011; Dembczyński *et al.*, 2010; Tsoumakas and Katakis, 2007]. However, with the increase of the number of labels, these standard multi-label classification methods that work in the original label space can easily become computationally impractical in training.

Recently, a number of label space reduction methods have been developed in the literature to address multi-label classification with many labels [Balasubramanian and Lebanon, 2012; Bi and Kwok, 2013; Hsu *et al.*, 2009; Chen and Lin, 2012; Tai and Lin, 2010; Zhou *et al.*, 2012]. These methods transform label vectors from high dimensional spaces to low dimensional spaces with random projections [Hsu *et al.*, 2009], maximum eigenvalue projections [Tai and Lin, 2010; Chen and Lin, 2012], Gaussian random projections [Zhou *et al.*, 2012], and label subset selections [Balasubramanian and Lebanon, 2012; Bi and Kwok, 2013], and then solve a small number of independent regression or classification tasks efficiently in the reduced output space. In addition to addressing multi-label classification with many labels, these label space reduction methods also share similar advantages as the feature space reduction methods on reducing the computational cost of training without much loss of prediction performance. However, all these current methods are limited to linear label space transformations and fail to capture nonlinear correlations between the multiple labels in the original label space.

In this paper, we first propose to exploit kernel canonical correlation analysis (KCCA) to perform feature-aware nonlinear label space dimension reduction for multi-label classification problems with many labels. Then we develop a unified approach that integrates both linear and nonlinear label space reductions based on canonical correlation analysis (CCA) and kernel canonical correlation analysis (KCCA) respectively to capture different types of label correlation patterns in the original label space. In particular, we employ a degree-2 polynomial nonlinear kernel in KCCA, which permits an efficient gradient descent decoding procedure in the test phase. After label space reduction, we solve a small number of independent regression problems in the transformed label space. The proposed approach works in a feature-aware manner since the label space reductions are conducted by using the feature inputs as a parallel view of the label vectors under the CCA and KCCA frameworks, which promotes the label predictability from the input features in the transformed label space. We conduct experiments on a number of multi-label classification datasets and the proposed approach demonstrates superior performance over a few state-of-the-art comparison methods.

2 Related Work

A significant number of multi-label learning works have been developed in the literature, most of which have centered on exploiting the interdependency information between labels, including probabilistic classifier chains [Dembczyński *et al.*, 2010; 2012; Kumar *et al.*, 2013], graphical model based methods [Guo and Gu, 2011; Ghamrawi and Maccallum, 2005], structured support vector machines [Tsochantaridis *et al.*, 2005] and max-margin methods [Guo and Schuurmans, 2011; Lampert, 2011].

The probabilistic classifier chains (PCC) method [Dembczyński *et al.*, 2010] applies the product rule of probability to the joint distribution of labels to capture conditional label dependencies. It estimates the conditional probability of every possible label set for an input instance and employs a Bayes optimal inference rule to optimize the given task loss function. The PCC method has appealing properties, but its accuracy is sensitive to the pre-specified ordering of the labels in training. Moreover, as suggested in [Dembczyński *et al.*, 2010], the applicability of the PCC method is limited to data sets with a small number of labels (no more than about 15 labels) and it is computationally intractable for problems with many labels. Some later works have tried to improve PCC by using an enhanced inference procedure [Dembczyński *et al.*, 2012] or applying beam search [Kumar *et al.*, 2013]. However the applicabilities of these improved methods are still limited to data sets with moderate number of labels. Ghamrawi and Maccallum [2005] explored conditional random field (CRF) classification models to parameterize label co-occurrences for multi-label classification. This work however suffers from the intractability of the exact inference problem for both training and testing processes due to the high tree-width graphical structures, while approximate inference methods converge to local optima. Its applicability has been limited to data sets with a small number of labels. The work of [Guo and Gu, 2011] uses a conditional dependency network to capture label dependencies, which also has similar computational problems and has only been applied on problems with small numbers of labels. The structured support vector machines (SSVMs) [Tsochantaridis *et al.*, 2005] have been developed to address prediction problems with structured and interdependent output variables. When applying SSVMs on multi-label classification problems with pairwise label-dependency structures, expensive inference procedures are involved during both training and testing phases. Some more recent max-margin works have developed novel loss functions [Guo and Schuurmans, 2011] and new formulations [Lampert, 2011] to specifically address multi-label classification. These methods however have only been applied on data sets with small numbers of labels. When the number of labels is large, these standard multi-label classification methods that work in the original label space can easily become computationally impractical.

Recent works on multi-label classification with many labels have focused on reducing the original large number of labels to a manageable set of transformed labels, by using linear label space projections and linear label subset selections [Hsu *et al.*, 2009; Chen and Lin, 2012; Lin *et al.*, 2014; Tai and Lin, 2010; Zhou *et al.*, 2012; Balasubramanian and Lebanon, 2012; Bi and Kwok, 2013].

An early work in [Hsu *et al.*, 2009] establishes a label projection framework to address multi-label classification with many labels. It first projects the high dimensional label vectors to a low dimensional space using a random transformation matrix, and then learns a regression model for each dimension of the transformed label vector. For a test instance, the estimated label vector from the regression models is then projected from the low dimensional space back to the original high dimensional label space. Following this framework, a number of improvements have been proposed. Tai and Lin [2010] proposed a principal label space transformation (PLST) method, which employs the principal component analysis (PCA) to reduce the label matrix in the original high dimensional space to a low dimensional space. Unlike random projections, the PCA dimensionality reduction produces the low dimensional representation by minimizing an L2-norm encoding error between the projected label matrix and the original label matrix. Subsequently, Chen and Lin [2012] proposed a conditional principal label space transformation (CPLST) method. It is a feature-aware method, which simultaneously minimizes both the L2-norm label encoding error and the least squares linear regression error in the reduced label space. Zhou *et al.* [2012] proposed a Gaussian random projection method for label space transformation. Recently, Lin *et al.* [2014] proposed a feature-aware implicit label space encoding (FaIE) method, which directly learns a latent code matrix and a linear decoding matrix by jointly maximizing the recoverability of the original label space and the predictability of the latent space.

In addition to these continuous label space dimensionality reductions, label subset selection methods directly select a discrete subset of the original labels to use. Balasubramanian and Lebanon [2012] proposed a multiple output prediction landmark selection method for multi-label classification with many labels. It selects a subset of the labels by minimizing the sparsity regularized encoding error. The approach in [Bi and Kwok, 2013] selects a small subset of the class labels from the original label space via randomized sampling.

These two groups of methods however are all limited to linear label space reductions, which fail to capture nonlinear label correlations in the original label space and may lead to severe information loss. Moreover, all these methods, except the works in [Chen and Lin, 2012; Lin *et al.*, 2014], perform label transformation on the label matrix in an "*unsupervised*" manner without taking the input feature information into consideration. This may produce transformed labels that are not well predictable from the input features. In this paper, we develop a *feature-aware nonlinear* label space reduction method for multi-label classification to address these two fundamental issues of existing works.

3 Preliminaries

In this section, we review the preliminaries over canonical correlation analysis (CCA) and kernel canonical correlation analysis (KCCA).

3.1 Canonical Correlation Analysis

Canonical correlation analysis (CCA) [Hotelling, 1936] is a well known tool for modeling linear associations between two sets of multi-dimensional variables. Given two views of the same set of objects, CCA projects each view into a low dimensional representation such that the two views are maximally correlated in the dimensionality reduced space. Traditionally, CCA has been used for supervised feature space dimensionality reduction in multi-label classification [Sun *et al.*, 2011], in which it treats the input features and the class labels as two parallel views of the same set of objects and projects the input data into a low dimensional space directed by the label information. Specifically, given an observed input data matrix $X \in \mathbb{R}^{t\times d}$, and its corresponding label indicator matrix $Y \in \{0,1\}^{t\times k}$, CCA projects them into a low dimensional space such that their correlation coefficient can be maximized. This can be formulated equivalently as the maximization problem below

$$\max_{\mathbf{u}\in\mathbb{R}^d,\mathbf{v}\in\mathbb{R}^k} \quad \mathbf{u}^\top X^\top Y\mathbf{v} \tag{1}$$
$$\text{subject to} \quad \mathbf{u}^\top X^\top X\mathbf{u} = 1, \quad \mathbf{v}^\top Y^\top Y\mathbf{v} = 1$$

which induces the following generalized eigenproblems on $\mathbf{u}$ and $\mathbf{v}$ [Hardoon *et al.*, 2004]

$$X^\top Y(Y^\top Y)^{-1}Y^\top X\mathbf{u} = \lambda^2 X^\top X\mathbf{u} \tag{2}$$
$$Y^\top X(X^\top X)^{-1}X^\top Y\mathbf{v} = \lambda^2 Y^\top Y\mathbf{v} \tag{3}$$

By solving these eigenproblems for the top m eigenvectors, CCA will find m pairs of projection vectors $\{(\mathbf{u}_i,\mathbf{v}_i)\}_{i=1}^m$, where $m < \min(k,d)$. By using all the $\{\mathbf{u}_i\}_{i=1}^m$ vectors as columns, one can form a $d\times m$ projection matrix U to perform linear dimensionality reduction over the input matrix X. Different from the unsupervised dimensionality reduction method PCA, whose orthogonal matrix is derived solely from the input data matrix, the orthogonal projection matrix U in CCA has the advantage of encoding label information from the label matrix Y.

Recently, CCA has also been considered for relating inputs to label projections in multi-label classification [Zhang and Schneider, 2011], where CCA is used to produce transformed labels to augment the original labels and increase label dimensions. By using all the $\{\mathbf{v}_i\}_{i=1}^m$ vectors as columns, one can form a $k\times m$ projection matrix V to perform linear dimensionality reduction over the label matrix Y. We can denote the CCA process that produces V as

$$V \leftarrow CCA(X,Y,m) \tag{4}$$

Using this projection matrix V produced under the guidance of the input data X, the label matrix Y can be mapped into a low dimensional $t\times m$ matrix $Z = YV$.

3.2 Kernel Canonical Correlation Analysis

The capacity of standard canonical correlation analysis (CCA) for data analysis and dimensionality reduction is limited by its linearity. To increase its capacity, kernel techniques have been used in CCA to produce a nonlinear extension, kernel canonical correlation analysis (KCCA) [Hardoon *et al.*, 2004].

To exploit kernel techniques, the original two view representations X and Y can be first mapped into high-dimensional feature spaces, $\Phi(X)$ and $\Psi(Y)$, respectively. Then the input kernel matrix can be obtained as $K_x = \kappa_x(X,X) = \Phi(X)\Phi(X)^\top$ and the label kernel matrix can be obtained as $K_y = \kappa_y(Y,Y) = \Psi(Y)\Psi(Y)^\top$, where $\kappa_x(\cdot,\cdot)$ and $\kappa_y(\cdot,\cdot)$ denote the kernel functions. Note that the high-dimensional representations, $\Phi(X)$ and $\Psi(Y)$, do not need to be given explicitly and one only needs to provide the kernel functions $\kappa_x(\cdot,\cdot)$ and $\kappa_y(\cdot,\cdot)$. With the input and label kernel matrices, KCCA maximizes the kernelized correlation coefficient, which can be equivalently formulated as the following maximization problem

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^t,\boldsymbol{\beta}\in\mathbb{R}^t} \quad \boldsymbol{\alpha}^\top K_x K_y\boldsymbol{\beta} \tag{5}$$
$$\text{subject to} \quad \boldsymbol{\alpha}^\top K_x K_x\boldsymbol{\alpha} = 1, \quad \boldsymbol{\beta}^\top K_y K_y\boldsymbol{\beta} = 1$$

Moreover, to cope with the potential singular problem of the kernel matrices, a regularization term ηI with a small $\eta > 0$ can be added to the constraints.[1] The regularized KCCA leads to the following generalized eigenproblems

$$K_x K_y\boldsymbol{\beta} = \lambda(K_x^2 + \eta I)\boldsymbol{\alpha} \tag{6}$$
$$K_y K_x\boldsymbol{\alpha} = \lambda(K_y^2 + \eta I)\boldsymbol{\beta} \tag{7}$$

which is equivalent to the unified eigenproblem below

$$\begin{pmatrix} 0 & K_xK_y \\ K_yK_x & 0 \end{pmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix} = \lambda\begin{pmatrix} K_x^2+\eta I & 0 \\ 0 & K_y^2+\eta I\end{pmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{bmatrix} \tag{8}$$

By solving this eigenproblem for the top q eigenvectors, KCCA will produce q pairs of projection vectors $\{(\boldsymbol{\alpha}_i,\boldsymbol{\beta}_i)\}_{i=1}^q$. One can then produce a projection matrix A for the input kernel matrix by using all $\{\boldsymbol{\alpha}_i\}_{i=1}^q$ vectors as columns of A, and produce a projection matrix B for the label kernel matrix by using all $\{\boldsymbol{\beta}_i\}_{i=1}^q$ vectors as columns of B. For the convenience of presentation, we denote the process of producing the B matrix as

$$B \leftarrow KCCA(K_x,K_y,q) \tag{9}$$

4 Nonlinear Label Space Transformation

In this section, we present a nonlinear label space transformation method to integrate both CCA and KCCA for multi-label classification with a large number of labels. The proposed methodology is motivated from two aspects: First, the input feature information guided label space dimensionality reduction can produce transformed labels that are easily predictable from the input features. Second, a combination of linear and nonlinear label transformations can capture both

[1] Similar regularization can be employed in the linear CCA case as well to ensure valid matrix inversions.

Algorithm 1 A Unified Training Algorithm

```
Input: $X$, kernel matrix $K_x$, label matrix $Y$,
    output kernel matrix $K_y$, $m$, $q$.
Output: label projection matrices $V$ and $B$,
    regression functions $h(\cdot)$ and $f(\cdot)$.
Procedure:
    1. perform CCA and KCCA:
       $V \leftarrow \text{CCA}(X, Y, m)$, $B \leftarrow \text{KCCA}(K_x, K_y, q)$.
    2. label transformation: $Z = YV$, $Q = K_yB$.
    3. learn a multi-dimensional regressor $h(\cdot)$ from
       the labeled data matrices $(X, Z)$.
    4. learn a multi-dimensional regressor $f(\cdot)$ from
       the labeled data matrices $(X, Q)$.
```

linear and nonlinear label correlations in the high dimensional label space and hence maximally reserve the original label information in the reduced label space. By integrating contributions from these two aspects, we expect the unified label space transformation method can effectively exploit potential label correlations to improve the high-dimensional multi-label classification performance.

Following the classic label projection framework established in the literature, our proposed approach has three steps: label encoding, independent regression, and label decoding. The first two steps form the training process and the decoding is performed to classify test instances.

4.1 A Unified Training Algorithm

Given a training data matrix $X \in \mathbb{R}^{t\times d}$, and its corresponding label indicator matrix $Y \in \{0,1\}^{t\times k}$, we first perform both linear and nonlinear feature-aware label space transformations to produce transformed label matrices in lower dimensional spaces. CCA can be directly performed over the label matrix Y and the input feature matrix X, using the process denoted by Equation (4), which produces a label space projection matrix V. Then the high dimensional label matrix Y can be projected into a low dimensional matrix $Z \in \mathbb{R}^{t\times m}$ by $Z = YV$.

KCCA is used to perform nonlinear label space transformation with the input data kernel matrix K_x and label kernel matrix K_y, aiming to capture nonlinear label correlations in the original label space. To enable an efficient decoding process from the reduced label vector to the original high dimensional label vector later in the test phase, in this work, we consider a polynomial kernel function with degree 2 as the label kernel function, $\kappa_y(\mathbf{y}, \mathbf{y}') = (\mathbf{y}^\top \mathbf{y}')^2$, such that the label kernel matrix is computed as

$$K_y = \kappa_y(Y, Y) = (YY^\top) \circ (YY^\top) \tag{10}$$

where $\circ$ denotes matrix Hadamard product. This polynomial kernel function can capture any pairwise dependence among the multiple labels. The input kernel matrix K_x can be computed using any mercer kernel functions. In our experiments later, we used a linear input kernel such that $K_x = XX^\top$. After performing KCCA over K_x and K_y, as indicated by Equation (9), to produce a label projection matrix B, the label kernel matrix can be projected into a lower dimensional matrix $Q \in \mathbb{R}^{t\times q}$ by $Q = K_yB$.

Algorithm 2 Decoding Algorithm

```
Input: a test instance $\mathbf{x}$, projection matrices $V$, $B$,
    regressor functions $h(\cdot)$, $f(\cdot)$,
    trade-off parameters $\mu \geq 0$, $0 \leq \gamma \leq 1$
Output: a solution label vector $\widehat{\mathbf{y}}^*$.
Procedure:
    set $\mathbf{z} = h(\mathbf{x})$, $\mathbf{q} = f(\mathbf{x})$
    initialize $\mathbf{y}^{(1)}$ as a $(k \times 1)$ zero vector
    for $r = 1, \ldots$ ,maxiters
        1. compute the gradient $g(\mathbf{y}^{(r)})$
        2. find the optimal step-size $\tau^*$
           using back-tracking line search in (19)
        3. set $\mathbf{y}^{(r+1)} = P_{\mathcal{C}}\big(\mathbf{y}^{(r)} - \tau^* \mathbf{g}(\mathbf{y}^{(r)})\big)$
        4. check convergence, break out if converged
    end for
    set $\widehat{\mathbf{y}}^* = \mathbf{y}^{(r+1)}$.
```

By combining both CCA and KCCA, we can obtain a unified projected label matrix $\widehat{Y} = [Z, Q]$, where each transformed label vector $\widehat{Y}_i$ is located in a low dimensional space $\mathbb{R}^{m+q}$ with $m + q < k$.

Next we learn a set of independent regression models from the input features to each dimension of the projected label matrix. This leads to training a multi-dimensional regressor on the *unified* transformed training data $(X, \widehat{Y})$, which is equivalent to training two multi-dimensional regressors $h(\cdot)$ and $f(\cdot)$ from the transformed training data (X, Z) and (X, Q) respectively.[2] Specifically, we conduct linear regression by minimizing the following regularized least squares losses:

$$\min_{W,\mathbf{b}} \quad \|Z - XW - \mathbf{1}\mathbf{b}^\top\|_F^2 + \alpha\|W\|_F^2 \tag{11}$$

$$\min_{\Omega,\mathbf{d}} \quad \|Q - X\Omega - \mathbf{1}\mathbf{d}^\top\|_F^2 + \alpha\|\Omega\|_F^2 \tag{12}$$

where $\|\cdot\|_F$ denotes the Frobenius matrix norm, $\mathbf{1}$ denotes a column vector with all 1 values. Closed-form solutions can be derived to solve these least squares regression problems efficiently in the reduced label spaces. Then the regressors $h(\cdot)$ and $f(\cdot)$ can be obtained using the trained model parameters

$$h(\mathbf{x}) = \mathbf{x}^\top W + \mathbf{b}^\top \tag{13}$$

$$f(\mathbf{x}) = \mathbf{x}^\top \Omega + \mathbf{d}^\top \tag{14}$$

The overall training algorithm is given in Algorithm 1.

4.2 Decoding

In the test phase, given a new test instance $\mathbf{x} \in \mathbb{R}^d$, we predict its label vector $\mathbf{y}$ in the original label space by solving a decoding problem. First, we can compute its regression label vectors, $\mathbf{z} \in \mathbb{R}^m$ and $\mathbf{q} \in \mathbb{R}^q$, in the reduced label spaces using the regressors trained above such that $\mathbf{z} = h(\mathbf{x})$ and $\mathbf{q} = f(\mathbf{x})$. Given the two label vectors, $\mathbf{z}$ and $\mathbf{q}$, in the reduced linear and nonlinear label spaces respectively, we next recover a unified $k \times 1$ label vector $\mathbf{y}$ in the original label

[2] For the convenience of algorithm presentation in the test phase, we use two multi-dimensional regressors instead.

space by minimizing a sparsity regularized joint least square loss function over both the linear and nonlinear label space transformations

$$\min_{\mathbf{y}} \quad \mu\|\mathbf{y}\|_1 + \frac{\gamma}{2}\|\mathbf{z} - V^\top\mathbf{y}\|^2 + \frac{1-\gamma}{2}\|\mathbf{q} - B^\top\kappa_y(Y, \mathbf{y})\|^2$$

$$\text{subject to} \quad \mathbf{y} \in \{0,1\}^k \tag{15}$$

where μ and γ are trade-off parameters. The $L1$-norm, $\|\mathbf{y}\|_1$, is used to promote the sparsity of the label vector. This is based on the observation that in many real world problems each instance is typically only assigned a few positive labels though the overall number of labels is large. The γ parameter balances the contributions of linear CCA and nonlinear KCCA label transformations for label vector recovery. For the polynomial label kernel function with degree 2, we have $\kappa_y(Y, \mathbf{y}) = (Y\mathbf{y}) \circ (Y\mathbf{y})$. This optimization problem however is hard to optimize due to the integer constraints over $\mathbf{y}$. We thus relax the integer constraints into linear inequality constraints $0 \leq \mathbf{y} \leq 1$, and solve the following relaxed optimization problem

$$\min_{\mathbf{y}} \quad \ell(\mathbf{y}) \qquad \text{subject to} \quad 0 \leq \mathbf{y} \leq 1 \tag{16}$$

where

$$\ell(\mathbf{y}) = \mu\|\mathbf{y}\|_1 + \frac{\gamma}{2}\|\mathbf{z} - V^\top\mathbf{y}\|^2 + \frac{1-\gamma}{2}\|\mathbf{q} - B^\top((Y\mathbf{y}) \circ (Y\mathbf{y}))\|^2 \tag{17}$$

We develop an efficient iterative projected gradient descent algorithm to solve the relaxed decoding minimization problem above. In the r-th iteration, the gradient of the objective function of (16) at the current point $\mathbf{y}^{(r)}$ can be computed as

$$\mathbf{g}(\mathbf{y}^{(r)}) \;=\; \mu\mathbf{1} + \gamma V(V^\top\mathbf{y}^{(r)} - \mathbf{z}) + \tag{18}$$
$$2(1-\gamma)Y^\top\Big[\Big(BB^\top\mathrm{diag}(Y\mathbf{y}^{(r)}\mathbf{y}^{(r)\top}Y^\top) - B\mathbf{q}\Big) \circ (Y\mathbf{y}^{(r)})\Big]$$

The next point $\mathbf{y}^{(r+1)}$ can then be reached by conducting a backtracking line search to find an optimal step-size τ^* and projecting the gradient update to the feasible set defined by the inequality constraints; that is,

$$\tau^* \;=\; \arg\min_{0\leq\tau\leq1} \; \ell\Big(P_{\mathcal{C}}\big(\mathbf{y}^{(r)} - \tau\mathbf{g}(\mathbf{y}^{(r)})\big)\Big) \tag{19}$$

$$\mathbf{y}^{(r+1)} \;=\; P_{\mathcal{C}}\big(\mathbf{y}^{(r)} - \tau^*\mathbf{g}(\mathbf{y}^{(r)})\big) \tag{20}$$

where $\mathcal{C} = \{\mathbf{y} \in \mathbb{R}^k : 0 \leq \mathbf{y} \leq 1\}$ is the feasible set of $\mathbf{y}$, and the projection function is defined as

$$P_{\mathcal{C}}(\cdot) = \min(\max(\cdot, 0), 1). \tag{21}$$

The overall decoding algorithm is given in Algorithm 2.

Finally, one can recover the $\{0,1\}$-valued label vector $\mathbf{y}^*$ of $\mathbf{x}$ in the original label space by rounding the solution $\widehat{\mathbf{y}}^*$ obtained from the relaxed optimization problem (16).

5 Experiments

To evaluate the proposed approach, we conducted experiments on five real-world multi-label datasets, comparing the proposed nonlinear approach to previous label reduction and transformation methods for multi-label classification with many labels. We report our experimental setting and results in this section.

Table 1: Statistical information of the datasets. Label card.: the average number of labels assigned to each instance.

Dataset	# of instances	# of labels	label card.
Corel5K	5,000	244	3.36
ESPGame	5,000	268	4.72
Iaprtc12	5,000	289	5.63
Enron	1,702	53	3.38
Delicious	16,105	983	19.02

5.1 Experimental Setting

We used five real world multi-label datasets for image and text categorization tasks in our experiments, including *Corel5K*, *ESPGame*, *Iaprtc12*, *Enron*, and *Delicious*. The first three datasets are image datasets. *Corel5K* is an important benchmark for keyword based image retrieval and annotation [Duygulu *et al.*, 2002]; *ESPGame* contains images obtained from an online game [von Ahn and Dabbish, 2004]; and *Iaprtc12* contains images initially published for cross-lingual retrieval [Makadia *et al.*, 2008]. From each of these three datasets, we constructed a subset to use by randomly sampling 5000 instances. *Enron* is a textual dataset that consists of 1702 email messages [Klimt and Yang, 2004]. The last dataset, *Delicious*, is a large scale textual dataset, which was extracted from a social bookmarking site and contains $16,105$ web pages along with 983 tags [Tsoumakas *et al.*, 2008]. Table 1 presents the statistical information of these datasets. We can see each of them has many label classes, and the average number of labels assigned to each instance is reasonably large, maintaining a valid multi-label classification problem with many labels in each case. For the image datasets, we used the GIST features [Oliva and Torralba, 2001], and each instance was represented as a 512-dimensional vector. We preprocessed each dataset by performing standardization over each column feature vector and normalization over each row instance vector.

In our experiments, we compared the following seven methods: (1) the proposed combination approach, denoted as *COMB*; (2) the nonlinear component of the combination approach, *KCCA*; (3) the linear component of the combination approach, *CCA*; (4) the Feature-aware Implicit Label space Encoding method (*FaIE*) [Lin *et al.*, 2014], which is a state-of-the-art label dimension reduction method for multi-label classification; (5) the label space transformation method [Zhang and Schneider, 2011], which constructs output codes using CCA to augment initial labels for multi-label classification and is denoted as *OC-CCA*; (6) the Conditional Principal Label Space Transformation method (*CPLST*) [Chen and Lin, 2012], which is another state-of-the-art label dimension reduction method for multi-label classification; and (7) the baseline method, partial binary relevance (*PBR*), from the empirical study of [Chen and Lin, 2012].

We conducted experiments using 10-fold cross validation on four datasets, except the large scale dataset *Delicious*, on which we conducted experiments using 5-fold cross validation. In each cross validation iteration, we performed parameter selection for all the comparison methods by us-

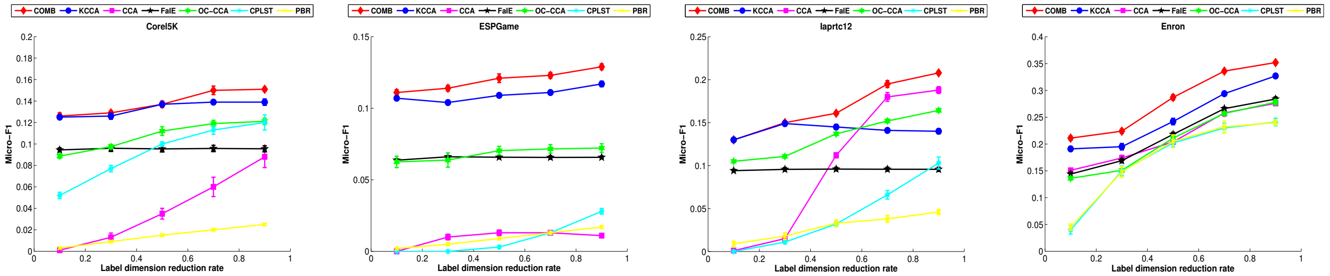

Figure 1: Comparison results in terms of Micro-F1 on all datasets with different label dimension reduction rate θ.

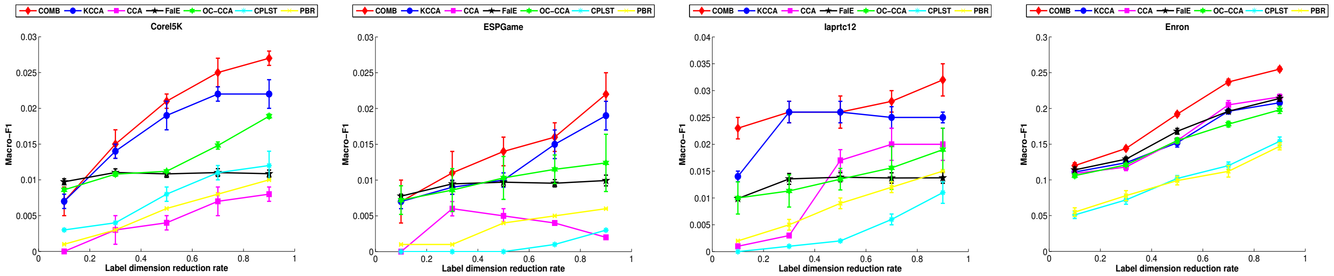

Figure 2: Comparison results in terms of Macro-F1 on all datasets with different label dimension reduction rate θ.

ing 80% of the training set for training and the remaining 20% for performance evaluation. For the proposed *COMB* method, there are two parameters μ and γ to be tuned for the decoding process. We selected the μ value from the set $[0.001, 0.005, 0.01, 0.05, 0.1]$, and selected the γ value from the set $[0, 0.2, 0.4, 0.6, 0.8, 1]$. Moreover, for the proposed *COMB* method, we need to split a given reduced label dimension $\widehat{k}$ into two parts, m and q, for its linear and nonlinear components. We used $m = round(\widehat{k}\gamma)$ and $q = \widehat{k} - m$ with the given trade-off parameter γ. For the four comparison methods from previous works, we performed parameter selection with values suggested in their original papers.

All the comparison approaches are evaluated using two popularly used multi-label classification evaluation measures, micro-F1 and macro-F1 [Tang *et al.*, 2009]. The micro-F1 measure gives equal weights to per-instance classification decision and is affected more by the major classes, whereas the macro-F1 measure gives equal weights to all classes and thus is more sensitive to the performance of rare classes.

5.2 Experimental Results

Let $\widehat{k}$ be the reduced label dimension for a given original label dimension k. We define $\theta = \widehat{k}/k$ as the *label dimension reduction rate*. Different θ values indicate different experimental settings: A smaller θ value indicates more severe label dimension reduction and possibly more information loss. Note for *OC-CCA*, we take the label dimension reduction rate of its CCA part as its θ value. For each dataset, we investigated a set of different label dimension reduction rates, $\theta \in [10\%, 30\%, 50\%, 70\%, 90\%]$. For each setting, we report the average and standard deviation results of the two F1 measures.

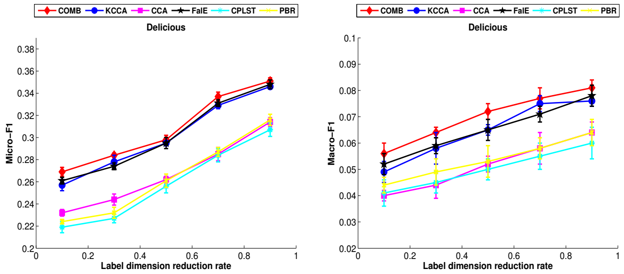

Figure 3: Comparison results in terms of Micro-F1 and Macro-F1 on *Delicious*.

Figure 1 presents the micro-F1 results of all the comparison methods on all the first four datasets, *Corel5K, ESPGame, Iaprtc12* and *Enron*, with different label dimension reduction rates. We can see that the three linear label space reduction methods, *CCA, FaIE* and *CPLST*, have strengths in different scenarios. For example, *FaIE* outperforms *CCA* and *CPLST* on *ESPGame* across different θ values; *CCA* outperforms *FaIE* and *CPLST* on *Iaprtc12* with large label reduction rates $\theta \in \{0.5, 0.7, 0.9\}$; and *CPLST* outperforms *FaIE* and *CCA* on *Corel5K* with large θ values $\{0.5, 0.7, 0.9\}$. They nevertheless demonstrate general advantages over the baseline *PBR* in most cases. But *FaIE* is the only one among the three that consistently outperforms *PBR* across all the datasets. The label transformation method *OC-CCA* also consistently outperforms the baseline *PBR* on all the datasets. *OC-CCA* also outperforms *FaIE* on *Iaprtc12* with different θ values and

Table 2: Running time (*training* and *testing*) on the five datasets (in seconds).

Methods	Corel5K		ESPGame		Iaprtc12		Enron		Delicious	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing
COMB	481	3006	537	4372	763	5266	154	708	3754	26687
KCCA	290	1857	374	2847	508	3173	110	463	2324	18937
CCA	267	1789	311	2631	412	2948	85	420	2143	17294
FaIE	563	296	634	336	709	358	175	82	2022	1504
CCA-OC	9973	556	13736	1053	13796	1526	2773	181	-	-
CPLST	8	4	9	6	12	5	3	2	44	9
PBR	7	3	8	5	11	4	2	1	40	8

on *Corel5K* with large θ values, while having similar performance with *FaIE* on *ESPGame* and *Enron*. The performance of *FaIE* however is quite stable across different label dimension reduction rates on the three image datasets. This suggests that most of the linear information in the label matrix of the image datasets can be captured with the low-dimensional orthonormal code matrix used by *FaIE*, while the nonlinear information can not be retrieved by increasing the code dimension. The nonlinear method *KCCA* significantly outperforms *CCA, FaIE, CPLST* and *OC-CCA* on the *Corel5K, ESPGame* and *Enron* datasets across different θ values. But it demonstrates inferior performance on the *Iaprtc12* dataset with large dimension reduction rates $\theta \in \{70\%, 90\%\}$. One possible reason is that *KCCA* takes all possible pairwise label correlations into account, which may capture some spontaneous noisy information under certain scenarios and induce overfitting. Hence smaller reduced label dimensions may lead to relatively more robust performance in such scenarios. The proposed combination method *COMB*, which integrates the strengths of *CCA* and *KCCA* in a *complementary* way, on the other hand demonstrates the strongest capacity of dealing with different datasets and different learning scenarios. It produces the best average results among all the comparison methods on all the four datasets across different label dimension reduction rates.

Similar results are presented in Figure 2 in terms of macro-F1 measure. *FaIE* demonstrates an overall better performance than the other two linear label dimension reduction methods, *CCA* and *CPLST*, in most scenarios. It also demonstrates consistent advantages over the baseline *PBR* across most cases. *OC-CCA* has similar but slightly better performance than *FaIE*. The nonlinear method *KCCA* again demonstrates better performance than the linear methods in most cases on three image datasets, *Corel5K, ESPGame, Iaprtc12*, but does not show much advantages on the *Enron* dataset. The proposed combination method *COMB* produces the best results among all the comparison methods across all the settings except only when $\theta = 0.1$ on *Corel5K* and *ESPGame*.

The comparison results in terms of the two F1 measures on the large scale dataset *Delicious* are presented in Figure 3. The *OC-CCA* method cannot handle this large scale dataset, we hence do not have results for this method. We can see that *FaIE* and the nonlinear *KCCA* perform much better than the other two methods *CCA* and *CPLST*, while *COMB* again produces the best average results across all cases.

In summary, all these results demonstrate that the proposed integrated nonlinear label space dimension reduction can better preserve the original label information across different scenarios. Moreover, we can see the performance of *COMB* is much better than both of its two components in most cases. This suggests that the proposed adaptive approach *COMB* is an effective integration model, instead of a simple switching procedure between its components, *CCA* and *KCCA*.

Running Time. To compare the empirical computational complexity of the comparison methods, we reported in Table 2 the training time and testing time of each method for a single run with θ=0.3 on a 64-bit PC with 4 processors (3.4 GHz) and 16 GB memory. We can see that *COMB* has longer running time than both *KCCA* and *CCA*, which is reasonable since *COMB* integrates the capacity of both *KCCA* and *CCA*. Nevertheless, *COMB* is more efficient than *FaIE* and *CCA-OC* in terms of training time on most datasets, except on *Iaprtc12* where *FaIE* is a bit faster and on *Delicious* where *FaIE* is faster but *CCA-OC* fails to run. Though *COMB* has higher testing time, it is still feasible to run on the large scale dataset. Moreover, the testing time can be significantly reduced if parallel resource is available, since test instances can be predicted independently. *CPLST* and *PBR* are efficient, but their poor performance cannot be compensated by time.

6 Conclusion

In this paper we proposed a novel nonlinear label space reduction method to address multi-label classification problems with high dimensional label spaces. The proposed approach integrates both linear CCA and nonlinear KCCA to perform label space dimensionality reduction in a feature-aware manner. It thus has the capacity of capturing both linear and nonlinear label correlation patterns in the original high dimensional label space and hence greatly increases the label information reservation level in the reduced label space. To recover a label vector in the original label space from the jointly reduced low dimensional space in the test phase, we formulated the decoding process as a sparsity regularized least square loss minimization problem and developed an efficient projected gradient descent algorithm to solve the minimization problem. We conducted experiments on a number of real-world multi-label datasets by comparing the proposed approach to a few state-of-the-art methods on multi-label classification with many labels. The empirical results demonstrated the efficacy of the proposed approach on capturing useful label information in the reduced label space and improving the performance of multi-label classification.

References

[Balasubramanian and Lebanon, 2012] K. Balasubramanian and G. Lebanon. The landmark selection method for multiple output prediction. In *Proceedings of ICML*, 2012.

[Bi and Kwok, 2013] W. Bi and J. Kwok. Efficient multi-label classification with many labels. In *Proceedings of ICML*, 2013.

[Chen and Lin, 2012] Y. Chen and H. Lin. Feature-aware label space dimension reduction for multi-label classification. In *Proceedings of NIPS*, 2012.

[Chen *et al.*, 2007] W. Chen, J. Yan, B. Zhang, Z. Chen, and Q. Yang. Document transformation for multi-label feature selection in text categorization. In *Proc. of ICDM*, 2007.

[Dembczyński *et al.*, 2010] K. Dembczyński, W. Cheng, and E. Hüllermeier. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proc. of ICML*, 2010.

[Dembczyński *et al.*, 2012] K. Dembczyński, W. Waegeman, and E. Hüllermeier. An analysis of chaining in multi-label classification. In *Proceedings of ECAI*, 2012.

[Duygulu *et al.*, 2002] P. Duygulu, K. Barnard, J. Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of ECCV*, 2002.

[Elisseeff and Weston, 2001] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *Proceedings of NIPS*, 2001.

[Ghamrawi and Maccallum, 2005] N. Ghamrawi and A. Maccallum. Collective multi-label classification. In *Proceedings of CIKM*, 2005.

[Guo and Gu, 2011] Y. Guo and S. Gu. Multi-label classification using conditional dependency networks. In *Proceedings of IJCAI*, 2011.

[Guo and Schuurmans, 2011] Y. Guo and D. Schuurmans. Adaptive large margin training for multilabel classification. In *Proceedings of AAAI*, 2011.

[Hardoon *et al.*, 2004] D. Hardoon, S. Szedmak, and J. Shawe-taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

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

[Hsu *et al.*, 2009] D. Hsu, S. Kakade, J. Langford, and T. Zhang. Multi-label prediction via compressed sensing. In *Proceedings of NIPS*, 2009.

[Klimt and Yang, 2004] B. Klimt and Y. Yang. The Enron corpus: A new dataset for email classification research. In *Proceedings of ECML*, 2004.

[Kumar *et al.*, 2013] A. Kumar, S. Vembu, A. Menon, and C. Elkan. Beam search algorithms for multi-label learning. *Machine Learning*, 92:65–89, 2013.

[Lampert, 2011] C. Lampert. Maximum margin multi-label structured prediction. In *Proceedings of NIPS*, 2011.

[Lewis *et al.*, 2004] D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *JMLR*, 5:361–397, 2004.

[Lin *et al.*, 2014] Z. Lin, G. Ding, M. Hu, and J. Wang. Multi-label classification via feature-aware implicit label space encoding. In *Proceedings of ICML*, 2014.

[Makadia *et al.*, 2008] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *Proceedings of ECCV*, 2008.

[Oliva and Torralba, 2001] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[Qi *et al.*, 2007] G. Qi, X. Hua, Y. Rui, J. Tang, T. Mei, and H. Zhang. Correlative multi-label video annotation. In *Proceedings of Multimedia*, 2007.

[Schapire and Singer, 2000] R. Schapire and Y. Singer. Boostexter: A boosting-based system for text categorization. *Machine Learning Journal*, pages 135–168, 2000.

[Sun *et al.*, 2011] L. Sun, S. Ji, and J. Ye. Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions, and analysis. *IEEE TPAMI*, 33:194–200, 2011.

[Tai and Lin, 2010] F. Tai and H. Lin. Multi-label classification with principal label space transformation. In *2nd Inter. Workshop on learning from Multi-Label Data*, 2010.

[Tang *et al.*, 2009] L. Tang, S. Rajan, and V. Narayanan. Large scale multi-label classification via metalabeler. In *Proceedings of Inter. WWW Conference*, 2009.

[Tsochantaridis *et al.*, 2005] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 6:1453–1484, 2005.

[Tsoumakas and Katakis, 2007] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *Inter. Journal of Data Warehousing and Mining*, 2007.

[Tsoumakas *et al.*, 2008] G. Tsoumakas, I. Katakis, and I. Vlahavas. Effective and efficient multilabel classification in domains with large number of labels. In *Proceedings of ECML/PKDD 2008 Workshop on MMD*, 2008.

[von Ahn and Dabbish, 2004] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI*, 2004.

[Zhang and Schneider, 2011] Y. Zhang and J. Schneider. Multi-label output codes using canonical correlation analysis. In *Proceedings of AISTATS*, 2011.

[Zhou and Zhang, 2006] Z. Zhou and M. Zhang. Multi-instance multi-label learning with application to scene classification. In *Proceedings of NIPS*, 2006.

[Zhou *et al.*, 2012] T. Zhou, D. Tao, and X. Wu. Compressed labeling on distilled lablsets for multi-label learning. *Machine Learning*, 88:69–126, 2012.