Multi-label Linear Discriminant Analysis

Hua Wang, Chris Ding, and Heng Huang

Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019, USA
huawang2007@mavs.uta.edu, chqding@uta.edu, heng@uta.edu

Abstract. Multi-label problems arise frequently in image and video annotations, and many other related applications such as multi-topic text categorization, music classification, *etc.* Like other computer vision tasks, multi-label image and video annotations also suffer from the difficulty of high dimensionality because images often have a large number of features. Linear discriminant analysis (LDA) is a well-known method for dimensionality reduction. However, the classical Linear Discriminant Analysis (LDA) only works for single-label multi-class classifications and cannot be directly applied to multi-label multi-class classifications. It is desirable to naturally generalize the classical LDA to multi-label formulations. At the same time, multi-label data present a new opportunity to improve classification accuracy through label correlations, which are absent in single-label data. In this work, we propose a novel Multi-label Linear Discriminant Analysis (MLDA) method to take advantage of label correlations and explore the powerful classification capability of the classical LDA to deal with multi-label multi-class problems. Extensive experimental evaluations on five public multi-label data sets demonstrate excellent performance of our method.

Keywords: Multi-label classification, Multi-label linear discriminant analysis, Image annotation.

1 Introduction

Image and video annotation has been an active research topic in recent years due to its potentially large impact on both image/video understanding and web/database image/video retrieval. In a typical image annotation problem, each picture is usually associated with several different conceptual classes. For example, the picture in Figure 1(a) is annotated with "building", "outdoor", and "urban", and similarly other pictures in Figure 1 are also associated with more than one semantic concepts. In machine learning, such problems that require each data point to be assigned to multiple different categories are called as *multi-label* classification problem. In contrast, in traditional *single-label* classification, also known as *single-label multi-class* classification, each data point belongs to only one category. Multi-label multi-class classification is more general than single-label multi-class classification, and recently has stimulated a slew of multi-label learning algorithms [16,5,7,10,3,9,17,4,15].

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part VI, LNCS 6316, pp. 126–139, 2010.

(a) building, outdoor, urban (b) face, person, entertainment (c) building, outdoor, urban (d) TV screen, person, studio

Fig. 1. Sample images from TRECVID 2005 data set. Each image is annotated with several different semantic words (listed under each images). When they are used as test images during cross-validations, our new proposed MLDA methods can correctly predict all of them. But other previous methods can only predict the first or second labels of each image. They cannot predict 'urban' in (a), 'entertainment' in (b), 'urban' in (c), 'person' and 'studio' in (d).

An important difference between single-label classification and multi-label classification lies in that, classes in the former are assumed mutually exclusive, while those in the latter are typically interdependent from one another. That is, in multi-label classification, class memberships can be inferred from one another through label correlations, which provide an important opportunity to improve classification accuracy. As a result, a multi-label classification method should make use of label correlations for improved classification performance.

High dimensionality of typical image data makes dimensionality reduction an important step to achieve efficient and effective image annotation. Among various dimensionality reduction methods in statistical learning, Linear Discriminant Analysis (LDA) is well known and widely used due to its powerful classification capability. However, LDA by nature is devised for single-label classification, therefore it can not be directly used in image annotation. The main difficulty to apply the classical LDA in multi-label classification is how to measure the inter and intra class scatters, which are clearly defined in single-label classification but become obscure in multi-label case. Because a data point with multiple labels belongs to different classes at the same time, how much it should contribute to the between-class and within-class scatters remains unclear. Therefore, it is desirable to generalize the classical LDA to deal with multi-label classification problem, and meanwhile, incorporate mutual correlations among labels.

In this paper, we propose a novel Multi-label Linear Discriminant Analysis (MLDA) method to explore the powerful classification capability of LDA in multi-label tasks and take advantage of label correlations. We first review the classical LDA and point out the computation ambiguity when using traditional single-label definitions of the scatter matrices in multi-label classification. After that, we introduce the details of our proposed MLDA method with empirical validations.

2 Difficulties of Classical LDA in Multi-Label Classification

Given a data set with n samples $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$ and K classes, where $\mathbf{x}_i \in \mathbb{R}^p$ and $\mathbf{y}_i \in \{0,1\}^K$. $\mathbf{y}_i(k) = 1$ if $\mathbf{x}_i$ belongs to the k-th class, and 0 otherwise. Let input data be partitioned into K groups as $\{\pi_k\}_{k=1}^K$, where π_k denotes the sample set of the k-th class with n_k data points. We write $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and

$$Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T = \left[\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(K)}\right], \tag{1}$$

where $\mathbf{y}_{(k)} \in \{0,1\}^n$ is the class-wise label indication vector for the k-th class.

2.1 Review of Classical LDA

Classical LDA seeks a linear transformation $G = \mathbb{R}^{p \times r}$ that maps $\mathbf{x}_i$ in the high p-dimensional space to a vector $\mathbf{q}_i \in \mathbb{R}^r$ in a lower $r(< p)$-dimensional space by $\mathbf{q}_i = G^T \mathbf{x}_i$. In classical LDA, the *between-class*, *within-class*, and *total-class* scatter matrices are defined as follows [2]:

$$S_b = \sum_{k=1}^{K} n_k \left(\mathbf{m}_k - \mathbf{m}\right)\left(\mathbf{m}_k - \mathbf{m}\right)^T, \tag{2}$$

$$S_w = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in \pi_k} \left(\mathbf{x}_i - \mathbf{m}_k\right)\left(\mathbf{x}_i - \mathbf{m}_k\right)^T, \tag{3}$$

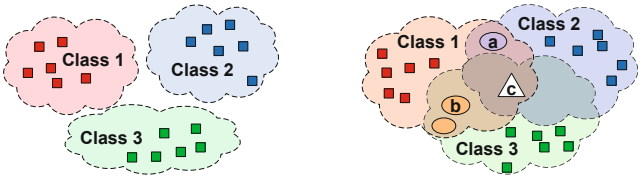
$$S_t = \sum_{i=1}^{n} \left(\mathbf{x}_i - \mathbf{m}\right)\left(\mathbf{x}_i - \mathbf{m}\right)^T, \tag{4}$$

where $\mathbf{m}_k = \frac{1}{n_k} \sum_{\mathbf{x}_i \in \pi_k} \mathbf{x}_i$ is the class mean (class centroid) of the k-th class, $\mathbf{m} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$ is the global mean (global centroid), and $S_t = S_b + S_w$. The optimal G is chosen such that the between-class distance is maximize whilst the within-class distance is minimized in the low-dimensional projected space, which leads to the standard LDA optimization objective [2] as follows:

$$\arg\max_G J = \mathbf{tr}\left(\frac{G^T S_b G}{G^T S_w G}\right). \tag{5}$$

2.2 Ambiguity Caused by Data with Multiple Labels in Classical LDA

Classical LDA for single-label classification is summarized in Eqs. (2–5), where the scatter matrices, S_w, S_b, and S_t, are well-defined as per the spatial distribution of data points as in Figure 2(a). However, in multi-label case, these definitions become obscure, because decision regions overlap among one another and decision boundaries turn out ambiguous as in Figure 2(b). Besides the data



(a) Single-label data. (b) Multi-label data.

Fig. 2. (a) In single-label classification, every data point distinctly belongs to only one class. (b) In multi-label classification, some data points may belong to multiple classes, denoted as ovals and triangles, which cause the ambiguity in scatter matrices calculations.

points belonging to only one class denoted by squares, some data points could also belong to multiple classes, as denoted by ovals for those belonging to two classes and triangles for those belonging to all three classes. How much a data point with multiple labels should contribute to the data scatters is not defined, therefore the scatter matrices defined in Eqs. (2–4) can not be computed.

3 Multi-label Linear Discriminant Analysis (MLDA)

Classical LDA deals with single-label problems, where data partitions are mutually exclusive, *i.e.*, $\pi_i \cap \pi_j = \varnothing$ if $i \neq j$, and $\sum_{k=1}^{K} n_k = n$. This, however, is no longer held in multi-label case. In this section, we propose a novel Multi-label Linear Discriminant Analysis (MLDA) method to explore the powerful classification capability of classical LDA in multi-label classification tasks. We first solve the ambiguity problem revealed in Section 2, and then leverage label correlations to enhance classification performance. Our method is a natural generalization of classical LDA.

3.1 Class-Wise Scatter Matrices

The ambiguity when using traditional single-label definitions of scatter matrices in multi-label classification prevents us from directly applying classical LDA to solve multi-label problems. Therefore, instead of defining the scatter matrices from data point perspective as in Eqs. (2–4), we propose to compute them by class-wise, such that the structural variances of training data are represented more lucidly and the scatter matrices are easier to be constructed. Moreover, the ambiguity, how much a data point with multiple labels should contribute to the scatter matrices, is avoided, and label correlations can be incorporated. The *class-wise between-class scatter matrix* is defined as:

$$S_b = \sum_{k=1}^{K} S_b^{(k)}, \;\; S_b^{(k)} = \left(\sum_{i=1}^{n} Y_{ik} \right) (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T, \tag{6}$$

the *class-wise within-class scatter matrix* S_w is defined as:

$$S_w = \sum_{k=1}^{K} S_w^{(k)},\ S_w^{(k)} = \sum_{i=1}^{n} Y_{ik} \left(\mathbf{x}_i - \mathbf{m}_k\right) \left(\mathbf{x}_i - \mathbf{m}_k\right)^T, \tag{7}$$

and the *class-wise total-class scatter matrix* S_t is defined as:

$$S_t = \sum_{k=1}^{K} S_t^{(k)},\ S_t^{(k)} = \sum_{i=1}^{n} Y_{ik} (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T, \tag{8}$$

where $\mathbf{m}_k$ is the mean of class k and $\mathbf{m}$ is the *multi-label global mean*, which are defined as follows:

$$\mathbf{m}_k = \frac{\sum_{i=1}^{n} Y_{ik}\mathbf{x}_i}{\sum_{i=1}^{n} Y_{ik}}, \quad \mathbf{m} = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n} Y_{ik}\mathbf{x}_i}{\sum_{k=1}^{K}\sum_{i=1}^{n} Y_{ik}}. \tag{9}$$

Theorem 1. *When applied into single-label classification, the multi-label scatter matrices, S_b, S_w, and S_t, defined in Eqs. (6–8), are reduced to their corresponding counterparts in classical LDA as defined in Eqs. (2–4).*

From the above definitions, the Theorem 1 can be easily obtained. Most importantly, in classical LDA, $S_t = S_b + S_w$, which is still held in multi-label classifications.

Theorem 2. *For multi-label class-wise scatter matrices, $S_b^{(k)}$, $S_w^{(k)}$, and $S_t^{(k)}$ as defined in Eqs. (6–8), the following relationship is held:*

$$S_t^{(k)} = S_b^{(k)} + S_w^{(k)}. \tag{10}$$

Therefore, $S_t = S_b + S_w$.

Proof. According to Eq. (9), we have $\sum_{i=1}^{n} Y_{ik}\mathbf{m}_k = \sum_{i=1}^{n} Y_{ik}\mathbf{x}_i$. Thus,

$$\sum_{i=1}^{n} Y_{ik}\mathbf{m}_k\mathbf{m}_k^T = \sum_{i=1}^{n} Y_{ik}\mathbf{m}_k\mathbf{x}_i^T \ \text{ and } \ \sum_{i=1}^{n} Y_{ik}\mathbf{m}_k\mathbf{m}_k^T = \sum_{i=1}^{n} Y_{ik}\mathbf{x}_i\mathbf{m}_k^T. \tag{11}$$

From Eqs. (6–8) and using Eq. (11), we have:

$$S_t^{(k)} = \sum_{i=1}^{n} Y_{ik}\mathbf{x}_i\mathbf{x}_i^T + \sum_{i=1}^{n} Y_{ik}\mathbf{m}\mathbf{m}^T - \sum_{i=1}^{n} Y_{ik}\mathbf{m}_k\mathbf{m}^T - \sum_{i=1}^{n} Y_{ik}\mathbf{m}\mathbf{m}_k^T = S_b^{(k)} + S_w^{(k)} \tag{12}$$

□

3.2 Multi-label Correlations

Multi-label data provide a new opportunity to improve classification accuracy through label correlations, which are absent in single-label data. Typically, the label correlation between two classes is formulated as following [15]:

$$C_{kl} = \cos(\mathbf{y}_{(k)}, \mathbf{y}_{(l)}) = \frac{\langle \mathbf{y}_{(k)}, \mathbf{y}_{(l)} \rangle}{\|\mathbf{y}_{(k)}\| \, \|\mathbf{y}_{(l)}\|}. \tag{13}$$

Thus, $C \in \mathbb{R}^{K \times K}$ is a symmetric matrix. Apparently, $C = I$ for single-label data. Namely, no label correlations can be utilized in single-label classification.

In multi-label classification, a data point may belong to several different classes simultaneously, hence the data points assigned to two different classes may overlap. Statistically, the bigger the overlap is, the more closely the two classes are related. Namely, class memberships in multi-label classification be inferred from one another through label correlations. Specifically, the *correlated labels assignments* are computed as:

$$Y^c = YC. \tag{14}$$

Several existing multi-label classification algorithms used label correlations to boost classification performance [16,1,3,15]. Using TRECVID 2005 data set with LSCOM-Lite annotation scheme [11], label correlations defined in Eq. (13) are illustrated in Figure 3. The high correlation value between "person" and "face" shows that they are highly correlated, which perfectly agree with the common sense in real life for the simplest fact that everybody has a face. Similar observations, such as "outdoor" and "sky", "waterscape-waterfront" and "boat-ship", "road" and "car", *etc.*, can also be seen in Figure 3, which concretely confirm the correctness of the formulation of label correlations defined in Eq. (13) from semantic perspective.

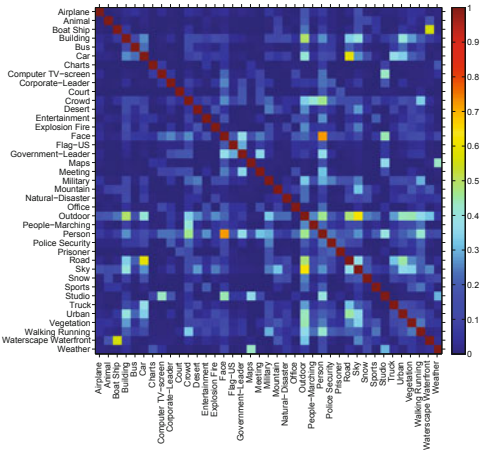


Fig. 3. Correlations between all pairs of 39 keywords in LSCOM-Lite on TRECVID 2005 data set

We replace Y by YC in Eqs. (6–9) in calculation of class-wise scatter matrices to incorporate label correlations. Theorems 1 still holds, because in single-label classification $C = I$ thereby $YC = Y$. Theorems 2 also holds, because we introduce C in both sides of equations.

3.3 Over-Counting Correction

Our further analysis on the class-wise scatter matrices in Eqs. (6–8) shows that the data points with multiple labels are over-counted in the scatter matrices calculations. For example, because data point **a** in Figure 2(b) has two labels for class 1 and class 2, it is used in both $S_b^{(1)}$ and $S_b^{(2)}$. Because $S_b = S_b^{(1)} + S_b^{(2)} + S_b^{(3)}$, data point **a** is used twice in the between-class scatter matrix S_b. Similarly, data point **c** is used three times in both S_b and S_w. In general, data point $\mathbf{x}_i$ with multiple labels is used $\sum_{k=1}^{K} \mathbf{y}_i(k)$ times in the scatter matrices, which are over-counted compared to data points associated with only one single label.

We correct the over-counting problem by introducing the normalized matrix $Z = [\mathbf{z}_1, \ldots, \mathbf{z}_n]^T \in \mathbb{R}^{n \times K}$:

$$\mathbf{z}_i = \mathbf{y}_i C / \|\mathbf{y}_i\|_{\ell 1}, \tag{15}$$

where $\|\cdot\|_{\ell 1}$ is the $\ell 1$-norm of a vector. A similar normalization could be as following:

$$\mathbf{z}_i = \mathbf{y}_i C / \|\mathbf{y}_i C\|_{\ell 1}, \tag{16}$$

such that $\sum_{k=1}^{K} \mathbf{z}_i(k) = 1$ and every data point has same importance in scatter matrices calculation. However, this is not reasonable for multi-label data when label correlations are considered, because a data point with multiple labels is generally believed to convey more information than that with only one single label. For example, in image annotation for natural scene pictures, a picture annotated with labels "Antarctica + penguin" is likely to contain more information than another one annotated with only label "Antarctic". Note that, $\sum_{k=1}^{K} \mathbf{z}_i(k) \geq 1$ when the *correlated normalized weight* in Eq. (15) is used, *i.e.*, the more labels a data point are associated with, the more important it is. Therefore, instead of using Eq. (16), in this work, we use the normalization in Eq. (15) to deal with the over-counting problem in multi-label data.

By replacing Y by Z in Eqs. (6–9), we have the final MLDA *scatter matrices*. Again, Theorems 1 and 2 can be similarly proved.

3.4 MLDA for Multi-label Classification

Now we write the scatter matrices in a more compact matrix form and summarize our MLDA method. First, let

$$\tilde{X} = X - \mathbf{m}\mathbf{e}^T, \tag{17}$$

where $\mathbf{e} = [1, \ldots, 1]^T$. Eq. (17) centers input data in multi-label sense, which is different from data centering in classical LDA for single-label classification where $\tilde{X} = X\left(I - \mathbf{e}\mathbf{e}^T/n\right)$.

We define $W = \mathrm{diag}(w_1, \ldots, w_K)$, where $w_k = \sum_{i=1}^{n} Z_{ik}$ is the weight of the k-th class in data scatters. Obviously, in single-label classification, $w_k = n_k$ is the number of data points in the k-th class. Thus,

$$S_b = \tilde{X} Z W^{-1} Z^T \tilde{X}^T. \tag{18}$$

Let $L = \text{diag}(l_1, \ldots, l_n)$, where $l_i = \sum_{i=1}^{K} Z_{ik}$. Clearly, in single-label classification, $L = I$, because each data point only belongs to one class. Thus,

$$S_t = \tilde{X} L \tilde{X}^T. \tag{19}$$

Finally, the optimization objective of our proposed MLDA method is defined in a similar way to classical LDA using trace of matrix ratio as following:

$$\arg\max_G \mathbf{tr}\left(\frac{G^T S_b G}{G^T S_w G}\right). \tag{20}$$

In real life applications, the number of features of a data set is often greater than that of training samples, thus S_w could be singular. As a result, in our implementation, we solve the eigenvalue problem $S_w^+ S_b \mathbf{v}_k = \lambda_k \mathbf{v}_k$, where S_w^+ is the pseudo-inverse of S_w. G is thus constructed by taking the eigenvectors corresponding to the r largest eigenvalues, and the classification can be carried out on the projected data.

4 Connections to Related Works

We review several most recent related multi-label classification methods which also use dimensionality reduction. First of all, many of these algorithms involve XYY^TX^T by certain forms in their optimization objectives, we thereby examine it in some details.

First, because $X\mathbf{y}_{(k)} = \sum_{\mathbf{x}_i \in \pi_k} Y_{ik}\mathbf{x}_i = w_k \mathbf{m}_k$, the following is held:

$$XYY^TX^T = \sum_{k=1}^{K} w_k^2 \mathbf{m}_k \mathbf{m}_k^T. \tag{21}$$

When the input data are properly centered as in Eq. (17), the between-class scatter matrix can be written as $S_b = \sum_{k=1}^{K} w_k \mathbf{m}_k \mathbf{m}_k^T$, thus XYY^TX^T is a coarse approximation of S_b. They are equivalent only if every class has same number of data points, *i.e.* $n_i = n_j$, $\forall i, j$.

Second, but more important, they treat the classes in a multi-label data set as independent, thereby label correlations, C, is not employed, though they are very important to enhance classification performance.

MLSI. Yu *et al.* [16] extended unsupervised latent semantic indexing (LSI) to make use of supervision information, called Multi-label informed Latent Semantic Indexing (MLSI) method using (in our notation)

$$\begin{aligned} \arg\max_G \mathbf{tr}\left(G^T\left((1-\beta)XX^TXX^T + \beta XYY^TX^T\right)G\right) \\ s.t. \quad G^TXX^TG = I. \end{aligned} \tag{22}$$

The first term is the original LSI objective. The second term is the supervised regularizer, which implicitly approximates S_b with deficiencies as analyzed above.

MDDM. Zhang *et al.* [17] proposed Multi-label Dimensionality reduction via Dependence Maximization (MDDM) method to identify a lower-dimensional

subspace by maximizing the dependence between the original features and associated class labels through (in our notation)

$$\max_G \ \mathbf{tr}\left(G^T X H Y Y^T H X^T G\right), \tag{23}$$

where $H = I - \mathbf{e}\mathbf{e}^T/n$ is the centering matrix in single-label sense such that XH has zero mean. However, the correct data centering in multi-label classification should be as in Eq. (17) and is different from XH. Ignoring H, Eq. (23) is same as Eq. (21), which simulates S_b without taking advantage of its full potentials.

MLLS. Ji *et al.* [3] suggested Multi-Label Least Square (MLLS) method to extract a common structure (subspace) shared among multiple labels. The optimization objective is (in our notation):

$$\max_G \mathbf{tr}\left(G^T \left(I - \alpha M\right)^{-1} \left(M^{-1} X Y Y^T X^T M^{-1}\right) G\right)$$
$$M = \frac{1}{n} X X^T + (\alpha + \beta) I. \tag{24}$$

Eq. (24) still fundamentally relies on XYY^TX^T to use label information, though more complicated.

We will compare the proposed MLDA method with these related algorithms in next evaluation section.

We notice another two recent works in [8] and [4] have close titles with our paper. However, the former attempts to solve multi-label classification implicitly through QR-decomposition in the null space of S_w, which is far more complicated than our method. Most importantly, label correlations are not considered in this work. The latter incorporates discriminative dimensionality reduction into Support Vector Machine (SVM), and thereby fundamentally different from the proposed MLDA method. In summary, our MLDA method present a generic framework for solving multi-label problems, which naturally incorporates label correlations inherent in multi-label data.

5 Experimental Results

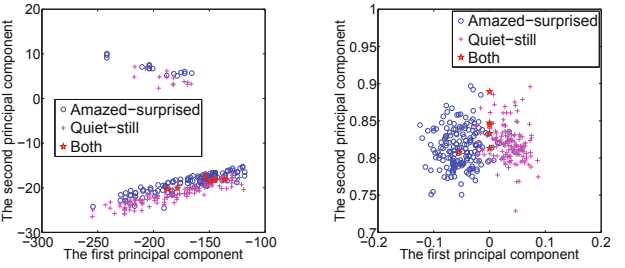
To evaluate the performance of multi-label classification methods, we use both basic image features (such as pixel values and moments of colors) and SIFT features in image classifications. We validate the proposed MLDA methods using the following standard multi-label data sets for image annotation.

TRECVID 2005[1] data set contains 61901 images and labeled with 39 concepts (labels). As most previous works, we randomly sample the data such that each concept has at least 100 images.

MSRC[2] data set is provided by the computer vision group at Microsoft Research Cambridge, which has 591 images annotated by 22 classes.

[1] http://www-nlpir.nist.gov/projects/trecvid

[2] http://research.microsoft.com/en-us/projects/objectclassrecognition/default.htm



(a) Visualization on 2D plane in the original space ($p = 72$).

(b) Visualization on 2D plane in the reduced subspace ($l = 5$) by MLDA.

Fig. 4. Visualization on 2D plane for the data points from the two classes in music emotion data set, in original space and projected space by MLDA, respectively

For these two image data sets, we divide each image into 64 blocks by a 8×8 grid and compute the first and second moments (mean and variance) of each color band to obtain a 384-dimensional vector as features. For MSRC data, we also employ SIFT features to measure similarities between images. We use MSRC (SIFT) to refer this data.

Mediamill [12] data set includes 43907 sub-shots with 101 classes, where each image is characterized by a 120-dimensional vector. Eliminating the classes containing less than 1000 samples, we have 27 classes. We randomly select 2609 sub-shots such that each class has at least 100 labeled data points.

In order to justify the generic applicability of our method, we also evaluate all methods on two following data sets from different applications.

Music emotion [13] data set comprises 593 songs with 6 emotions (labels). The dimensionality of the data points is 72.

Yahoo data set described in [14] came from the "yahoo.com" domain. Each web page is described as a 37187-dimensional feature vector. We use the "science" topic because it has maximum number of labels, which contains 6345 web pages with 22 labels.

5.1 Discriminative Capability of MLDA

We first evaluate the discriminative capability of the proposed MLDA method. We randomly pick up two classes from music emotion data set, "amazed-surprised" and "quiet-still", and visualize the data points from these two classes in the original space ($p = 72$) on 2D plane using the first two principal component coordinates as shown in Figure 4(a). It is obvious that the data points are mingled together and it is difficult to find a linear decision boundary with high classification accuracy. We then run MLDA on the whole data set with all six labels, and transform the data points into the obtained projection subspace ($l = K - 1 = 5$), in which we visualize the same data points on 2D plane as shown in Figure 4(b). Apparently, the

Table 1. Performance evaluations of six compared methods by 5-fold cross validations

Data	Evaluation metrics		Compared methods					
			LDA-C1	SVM	MLSI	MDDM	MLLS	MLDA
TREC05	Macro average	Precision	0.282	0.269	0.247	0.366	0.248	**0.420**
		F1 score	0.190	0.286	0.275	0.370	0.276	**0.399**
	Micro average	Precision	0.274	0.252	0.234	0.352	0.241	**0.418**
		F1 score	0.408	0.399	0.293	0.491	0.295	**0.528**
MSRC	Macro average	Precision	0.291	0.274	0.252	0.370	0.255	**0.431**
		F1 score	0.201	0.295	0.287	0.392	0.290	**0.410**
	Micro average	Precision	0.288	0.262	0.253	0.363	0.255	**0.420**
		F1 score	0.415	0.406	0.301	0.504	0.302	**0.533**
MediaMill	Macro average	Precision	0.337	0.302	0.207	0.385	0.206	**0.410**
		F1 score	0.349	0.322	0301	0.418	0.311	**0.430**
	Micro average	Precision	0.335	0.297	0.207	0.382	0.205	**0.388**
		F1 score	0.518	0.398	0.341	0.440	0.340	**0.443**
Music emotion	Macro average	Precision	0.507	0.434	0.329	0.509	0.311	**0.614**
		F1 score	0.453	0.418	0.323	0.506	0.471	**0.618**
	Micro average	Precision	0.504	0.501	0.328	0.507	0.308	**0.613**
		F1 score	0.477	0.441	0.339	0.518	0.475	**0.626**
Yahoo (Science)	Macro average	Precision	0.458	0.414	0.396	0.463	0.421	**0.501**
		F1 score	0.227	0.302	0.296	0.481	0.443	**0.498**
	Micro average	Precision	0.447	0.416	0.395	0.458	0.420	**0.499**
		F1 score	0.226	0.218	0.209	0.484	0.519	**0.544**
MSRC (SIFT)	Macro average	Precision	0.415	0.408	0.428	0.520	0.424	**0.612**
		F1 score	0.367	0.358	0.381	0.471	0.376	**0.531**
	Micro average	Precision	0.408	0.403	0.412	0.515	0.407	**0.597**
		F1 score	0.612	0.611	0.620	0.671	0.617	**0.698**

data points are clearly separated according to their class membership now, which demonstrates that the projection subspace produced by MLDA is indeed more discriminative. In addition, MLDA significantly reduces the data dimensionality (from 72 to 5), such that the computational complexity of the subsequent classification is largely reduced.

5.2 Classification Performance

We use standard 5-fold cross validation to evaluate the classification performance of the proposed MLDA method, and compare the results to the three related multi-label classification methods, MLSI, MDDM, and MLLS discussed in Section 4. K-Nearest Neighbor (KNN) classifier ($K = 1$ in this work) is used for classification after dimensionality reduction by MLSI, MDDM, and MLDA methods. We also

tested $K = 3, 5$ and the results are similar to $K = 1$. Because of the limited space, we only show the results of $K = 1$. Euclidean distance is used to decide neighborhood in KNN. KNN is conducted one class at a time, where a binary classification is conducted for each class. Note that, we choose KNN because it is the most widely used classification method following standard LDA. Because multi-label problem is already addressed in the dimensionality reduction step in MLSI, MDDM and our method, the subsequent classification method, such as KNN in our evaluations, do not need to take care of multi-label issue any longer. MLLS has its own classification mechanism. Following the standard way, we select $l = K - 1$ as the dimensionality of the projected subspace. For MLSI, the parameter β is set as 0.5 as recommended in [16]. For MDDM, we use the same linear kernel as in the experimental evaluation in [17]. For MLLS, we use the codes posted at the authors' web site [3], which fine tunes the parameters based on F1 scores.

LDA-C1. We report the classification performance of classical LDA as a reference. Because classical LDA is inherently a single-label classification method, we do both dimensionality reduction and classification one class at a time. For every class, the classification is done as a binary classification problem, which thereby implicitly treats all the classes isolated.

Support Vector Machine (SVM). We use SVM classification results as a baseline. Similar to LDA-C1, we run SVM one class at a time, and for every class the classification is done as a binary classification problem. SVM is implemented by LIBSVM[3] (Matlab version).

The conventional classification performance metrics in statistical learning, *precision* and *F1 score*, are used to evaluate the compared methods. Precision and F1 score are computed for every class following the standard definition for a binary classification problem. To address multi-label classification, class-wise macro average and micro average are used to assess the overall performance across multiple labels [6]. In multi-label classification, the macro average is the mean of the values of a standard class-wise metric over all the labels, thus attributing equal weights to every class. The micro average is obtained from the summation of contingency matrices for all binary classifiers. The micro average metric gives equal weight to all classifications, which can be seen as a weighted average that emphasizes more on the accuracy of categories with more positive samples.

Table 1 presents the classification performance comparisons by 5-fold cross validation, which show that the proposed MLDA method generally outperforms all other methods, sometimes significantly. We achieve about 10% improvement on average over all the data sets. To be more specific, Figure 1 shows four images from TRECVID 2005. Only the proposed MLDA method can correctly annotate all images. All other methods only predict part of the labels. These results quantitatively demonstrate the effectiveness of our method, and justify the utility of the class-wise scatter matrices and label correlations.

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

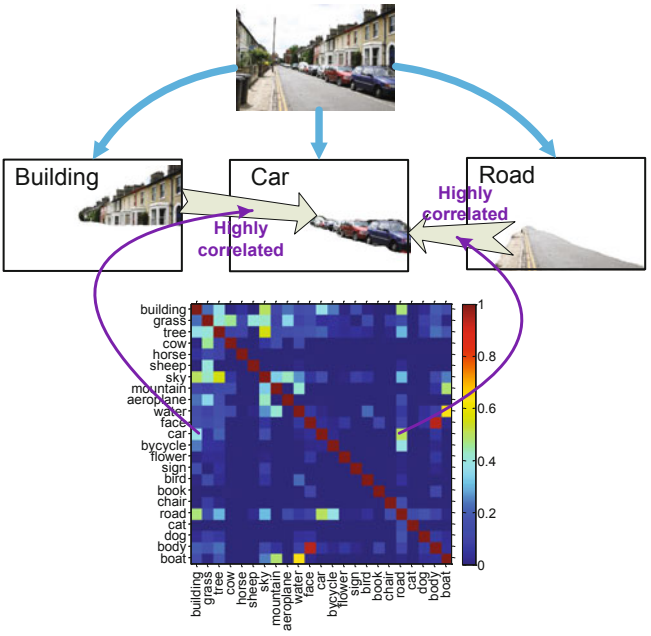


Fig. 5. An example image from MSRC data (Top). The label "car" can only be correctly annotated by our MLDA method, because "car" has high correlations with "building" and "road" (appearing in the image). The bottom panel visualizes the label correlation matrix.

5.3 Label Transfer via Label Correlations

A more careful examination on the classification results in Section 5.2 shows that, for the sample image shown in the top panel of Figure 5 from MSRC data set, the label "car" can only be correctly annotated by the proposed MLDA method, while two other labels, "building" and "road", generally can be correctly annotated by most of the compared methods. By scrutinizing label correlations of MSRC data set, defined by Eq. (13) and illustrated in the bottom panel of Figure 5, we can see that "car" is highly correlated with both "building" and "road". Therefore, label "car" is transferred to the sample image from its annotated labels through label correlations, which concretely corroborates the usefulness of label correlations to boost multi-label classification performance.

6 Conclusions

In this work, we proposed a novel Multi-label Linear Discriminant Analysis (MLDA) method to naturally generalize classical LDA for multi-label classification. We reformulated the scatter matrices from class perspective, such that

the new class-wise scatter matrices solved the computation ambiguity to use traditional single-label definitions of scatter matrices in multi-label classification, and incorporated label correlations from multi-label data. We examined three closely related multi-label classification methods and showed the advantages of our method theoretically. Encouraging results in extensive experimental evaluations supported our proposed methods and theoretical analysis empirically.

Acknowledgments. This research is supported by NSF-CCF 0830780, NSF-CCF 0939187, NSF-CCF 0917274, NSF-DMS 0915228, NSF-CNS 0923494.

References

1. Chen, G., Song, Y., Wang, F., Zhang, C.: Semi-supervised Multi-label Learning by Solving a Sylvester Equation. In: Proc. of SDM (2008)
2. Fukunaga, K.: Introduction to statistical pattern recognition (1990)
3. Ji, S., Tang, L., Yu, S., Ye, J.: Extracting shared subspace for multi-label classification. In: Proc. of SIGKDD, pp. 381–389 (2008)
4. Ji, S., Ye, J.: Linear Dimensionality Reduction for Multi-label Classification. In: Proc. of IJCAI, pp. 1077–1082 (2009)
5. Kang, F., Jin, R., Sukthankar, R.: Correlated label propagation with application to multi-label learning. In: Proc. of CVPR, pp. 1719–1726 (2006)
6. Lewis, D., Yang, Y., Rose, T., Li, F.: Rcv1: A new benchmark collection for text categorization research. Journal of Machine Learning Research 5, 361–397 (2004)
7. Liu, Y., Jin, R., Yang, L.: Semi-supervised multi-label learning by constrained non-negative matrix factorization. In: Proc. of AAAI, p. 421 (2006)
8. Park, C., Lee, M.: On applying linear discriminant analysis for multi-labeled problems. Pattern Recognition Letters 29(7), 878–887 (2008)
9. Pock, T., Schoenemann, T., Graber, G., Bischof, H., Cremers, D.: A convex formulation of continuous multi-label problems. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 792–805. Springer, Heidelberg (2008)
10. Qi, G., Hua, X., Rui, Y., Tang, J., Mei, T., Zhang, H.: Correlative multi-label video annotation. In: Proc. of ACM Multimedia, pp. 17–26 (2007)
11. Smeaton, A., Over, P., Kraaij, W.: Evaluation campaigns and TRECVid. In: Proc. of MIR, p. 330 (2006)
12. Snoek, C.G.M., Worring, M., van Gemert, J.C., Geusebroek, J.M., Smeulders, A.W.M.: The challenge problem for automated detection of 101 semantic concepts in multimedia. In: Proc. of ACM Multimedia, pp. 421–430 (2006)
13. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multilabel classification of music into emotions. In: Proc. of ISMIR (2008)
14. Ueda, N., Saito, K.: Single-shot detection of multiple categories of text using parametric mixture models. In: Proc. of SIGKDD, pp. 626–631 (2002)
15. Wang, H., Huang, H., Ding, C.: Image Annotation Using Multi-label Correlated Greens Function. In: Proc. of ICCV (2009)
16. Yu, K., Yu, S., Tresp, V.: Multi-label informed latent semantic indexing. In: Proc. of SIGIR, p. 265 (2005)
17. Zhang, Y., Zhou, Z.: Multi-Label Dimensionality Reduction via Dependence Maximization. In: Proc. of AAAI, pp. 1503–1505 (2008)