Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)

# Discriminant Laplacian Embedding

**Hua Wang, Heng Huang** and **Chris Ding**
Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76013
huawang2007@mavs.uta.edu, {chqding,heng}@uta.edu

## Abstract

Many real life applications brought by modern technologies often have multiple data sources, which are usually characterized by both attributes and pairwise similarities at the same time. For example in webpage ranking, a webpage is usually represented by a vector of term values, and meanwhile the internet linkages induce pairwise similarities among the webpages. Although both attributes and pairwise similarities are useful for class membership inference, many traditional embedding algorithms only deal with one type of input data. In order to make use of the both types of data simultaneously, in this work, we propose a novel Discriminant Laplacian Embedding (DLE) approach. Supervision information from training data are integrated into DLE to improve the discriminativity of the resulted embedding space. By solving the ambiguity problem in computing the scatter matrices caused by data points with multiple labels, we successfully extend the proposed DLE to multi-label classification. In addition, through incorporating the label correlations, the classification performance using multi-label DLE is further enhanced. Promising experimental results in extensive empirical evaluations have demonstrated the effectiveness of our approaches.

## Introduction

Embedding techniques seek to represent input data in their lower-dimensional "intrinsic" subspace/sub-manifold, in which irrelevant features are pruned and inherent data structures are more lucid. In the early ages, embedding algorithms assume that input data objects are homogeneous but not relational, and thereby are devised to deal with a set of *attributes* in the format of fixed-length vectors. For example, Principal Component Analysis (PCA) (Jolliffe 2002) attempts to maximize the covariance among data points, and Linear Discriminant Analysis (LDA) (Fukunaga 1990) aims at maximizing the class separability. In recent years, manifold learning motivates many embedding algorithms using *pairwise similarities* between data objects, such as ISOMAP (Tenenbaum, Silva, and Langford 2000), Locally Linear Embedding (LLE) (Roweis and Saul 2000), Laplacian Eigenmap (Belkin and Niyogi 2002) and Locality Preserving Projection (LPP) (He and Niyogi 2003), *etc*. These algorithms generally assume that the observed data are sampled from an underlying sub-manifold which is embedded in a high-dimensional observation space.



With the advances in science and technology, many data sets in real applications nowadays are no longer confined to one single format but come from multiple different sources. Thus, for a same data set, we often have both attributes information about data objects (vector data) and various pairwise similarities between data objects (graph data) simultaneously. For example, a webpage can be represented as a vector of term values. In addition, there also exist hyperlinks between webpages, which indicates a similarity relationship among the webpages. Because the both types of data, attributes and pairwise similarities, convey valuable information for classification, the aforementioned embedding algorithms designed for only one single type of data are usually insufficient, especially the correlations between these two types of data are not exploited. In this work, we propose an novel Discriminant Laplacian Embedding (DLE) approach to integrate these two types of data. Our new algorithm is interesting from a number of perspectives as following.

- Using an integral optimization objective, DLE integrates both attributes data and pairwise similarity data in a shared low-dimensional subspace, which is infused with the LDA space of vector data and the Laplacian eigenspace for graph data. As a result, the discriminability of data are incorporated in the both spaces.
- Unlike many existing embedding algorithms purely relying on unsupervised data, the labels of training data are utilized in DLE, such that the performance of classifications on the projected data via DLE is improved.
- DLE is linear and thereby fast, which makes it suitable for practical applications. Through DLE, new data points can be directly placed in the projection space. This is different from many existing nonlinear embedding algorithms, by which new data points must be interpolated in the representation space/manifold.
- Through a natural extension, the proposed DLE approach can be naturally extended to multi-label classification, an emerging topic in machine learning in recent years, where each objects may belongs to more than one class.

## Discriminant Laplacian Embedding

For a classification task with $n$ data points and $K$ classes, each data point $\mathbf{x}_i \in \mathbb{R}^p$ is associated with a subset of class

labels represented by a binary vector $\mathbf{y}_i \in \{0,1\}^K$ such that $\mathbf{y}_i(k) = 1$ if $\mathbf{x}_i$ belongs to the $k$th class, and 0 otherwise. In single-label classification, each data point belongs to only one class, $\sum_k \mathbf{y}_i(k) = 1$, while in multi-label classification, each data point may belong to multiple classes at the same time, $\sum_k \mathbf{y}_i(k) \geq 1$. Meanwhile, we also have the pairwise similarities $W \in \mathbb{R}^{n\times n}$ among the $n$ data points with $W_{ij}$ indicating how close $\mathbf{x}_i$ and $\mathbf{x}_j$ are related. Suppose the number of labeled data points is $l(< n)$, our goal is to predict labels $\{\mathbf{y}_i\}_{i=l+1}^n$ for the unlabeled data points $\{\mathbf{x}_i\}_{i=l+1}^n$. We write $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$.

Traditional embedding algorithms learn a projection only from one type of data, either attribute data $X$, such as PCA and LDA, or pairwise similarity data $W$, such as Laplacian Eigenmap and LPP. Besides, though useful, the label information from training data $\{\mathbf{y}_i\}_{i=1}^l$ is not always used, because many embedding algorithms are devised closely in conjunction with clustering algorithms which are unsupervised by nature. Therefore, we propose a novel Discriminant Laplacian Embedding (DLE) approach to realize these two expectations, data integration and making use of supervision information, which produces a transformation $U \in \mathbb{R}^{p\times r}$ by solving the following optimization objective:

$$\arg\max_U \mathbf{tr}\left(U^T\left(A_+^{-\frac{1}{2}} {S_w}^{-\frac{1}{2}} S_b {S_w}^{-\frac{1}{2}} A_+^{-\frac{1}{2}}\right)U\right), \quad (1)$$

where $r(\ll p)$ is the dimensionality of the projected subspace, $S_b$ and $S_w$ are the between-class and within-class scatter matrices defined same as in standard LDA, and $A = X\left(D - W\right)X^T$ with $D = \mathrm{diag}\left(d_1, \ldots, d_n\right), d_i = \sum_j W_{ij}$. Thus, $L = D - W$ is the graph Laplacian (Chung 1997). The solution to Eq. (1) is well established in mathematics by solving the following eigenvalue problem:

$$\left(A_+^{-\frac{1}{2}} {S_w}^{-\frac{1}{2}} S_b {S_w}^{-\frac{1}{2}} A_+^{-\frac{1}{2}}\right)\mathbf{u}_k = \lambda_k \mathbf{u}_k, \quad (2)$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ are the resulted eigenvalues and $\mathbf{u}_k$ are the corresponding eigenvectors. Hence, $U = [\mathbf{u}_1, \ldots, \mathbf{u}_r]$, and $r$ is empirically selected as $K - 1$.

Then we may project $\mathbf{x}_i$ into the embedding space as:

$$\mathbf{q}_i = U^T \mathbf{x}_i, \quad (3)$$

and subsequent classification is carried out using $\mathbf{q}_i$. In this work, we use $K$-nearest neighbor (KNN) method (Fukunaga 1990) for classification because of its simplicity and clear intuition ($K = 1$ is used in this work and we abbreviate it as 1NN). For each class, the classification is conducted as a binary classification problem, one class at a time.

In the rest of this section, we will derive the optimization objective in Eq. (1).

## Backgrounds of DLE

Given training data $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^l$, LDA targets on projecting the original data from a high $p$-dimensional space to a much lower $r$-dimensional subspace to separate different classes as much as possible (maximize between-class scatter $S_b$), whilst condense each class as much as possible (minimize within-class scatters $S_w$). Computing $S_b$ and $S_w$ by the standard definitions in LDA (Fukunaga 1990), we achieve the standard LDA optimization objective to maximize:

$$J_{\text{LDA}} = \mathbf{tr}\left(\frac{G^T S_b G}{G^T S_w G}\right). \quad (4)$$

The transformation $G \in \mathbb{R}^{p\times r}$ is usually obtained by applying the eigen-decomposition on ${S_w}^{-1} S_b$ (or sometimes by solving the generalized eigenvalue problem $S_b \mathbf{u}_k = \lambda_k S_w \mathbf{u}_k$), when $S_w$ is nonsingular. This method, however, can not guarantee the orthonormality of $G$. To this end, learning from the way frequently used in spectral clustering, let $F = {S_w}^{\frac{1}{2}} G$, we maximize:

$$J_{\text{LDA}} = \mathbf{tr}\left(F^T {S_w}^{-\frac{1}{2}} S_b {S_w}^{-\frac{1}{2}} F\right), \quad (5)$$

where the constraint $F^T F = I$ is automatically satisfied and thereby removed from the formulation.

Given the pairwise similarity $W$, Laplacian embedding preserves the same relationships and maximize the smoothness with respect to the intrinsic manifold of the data set in the embedding space by minimizing (Hall 1970)

$$\sum_{i,j} \|\mathbf{q}_i - \mathbf{q}_j\|^2 W_{ij} = \mathbf{q}_i^T \left(D - W\right)\mathbf{q}_i. \quad (6)$$

For multi-dimensional embedding, Eq. (6) becomes $\mathbf{tr}\left(Q^T(D - W)Q\right)$, where we use linear embedding $Q^T = [\mathbf{q}_1, \ldots, \mathbf{q}_n] = F^T X$. We hence write Eq. (6) as follows:

$$J_{\text{Lap}} = \mathbf{tr}\left(F^T X(D - W)X^T F\right). \quad (7)$$

## Motivations and Formulation of DLE

As we are seeking an embedding to leverage both attributes and pairwise similarity of a same data set, given two individual optimization objectives as in Eq. (5) and Eq. (7), we may construct an additive embedding objective to maximize:

$$J_{\text{DLE}} = \alpha J_{\text{LDA}} - (1 - \alpha) J_{\text{Lap}}, \quad (8)$$

where $0 < \alpha < 1$ is a tradeoff parameter. In practice, however, it is hard to choose an optimal $\alpha$. Therefore, instead of using the trace of *difference*, we formulate the objective as the trace of *quotient* so that $\alpha$ is removed as follows:

$$J_{\text{DLE}} = \mathbf{tr}\left(\frac{F^T {S_w}^{-\frac{1}{2}} S_b {S_w}^{-\frac{1}{2}} F}{F^T X(D - W)X^T F}\right). \quad (9)$$

Let $A = X(D - W)X^T$ and $U = A_+^{\frac{1}{2}} F$, we obtain the symmetric optimization objective of our proposed DLE as in Eq. (1), which yields orthonormal $U$ as solution.

Considering the fact that $A$ is usually rank deficient, let

$$A = V \begin{bmatrix} \Sigma & \\ & \mathbf{0} \end{bmatrix} V^T \quad (10)$$

be the eigen-decomposition of $A$ with diagonal line of $\Sigma$ being the positive eigenvalues of $A$, we have

$$A_+^{\frac{1}{2}} = V_1 \Sigma^{\frac{1}{2}} V_1^T \quad \text{and} \quad A_+^{-\frac{1}{2}} = V_1 \Sigma^{-\frac{1}{2}} V_1^T, \quad (11)$$

Table 1: Semi-supervised classification using DLE.

**Input**:
$\{\mathbf{x}_i\}_{i=1}^n$: Attribute vectors for all the data points.
$\{\mathbf{y}_i\}_{i=1}^l$: labels for the training data points.
$W$: pairwise relationship among all the data points.

**Steps**:
(1) Compute $S_b$ and $S_w$ as in standard LDA algorithm.
(2) Compute $A = X(D - W)X^T$.
(3) Resolve the eigenvalue problem in Eq. (2). Construct the projection matrix $U$ by the resulted eigenvectors corresponding to the $(K-1)$ largest eigenvalues.
(4) Compute the projected vectors $\{\mathbf{q}_i\}_{i=1}^n$ for all the data points including those unlabeled as in Eq. (3).
(5) Use projected training data $\{(\mathbf{q}_i, \mathbf{y}_i)\}_{i=1}^l$ to classify the testing data points $\{\mathbf{q}_i\}_{i=l+1}^n$ via 1NN method.

**Output**:
Class labels for testing data points $\{\mathbf{y}_i\}_{i=l+1}^n$.

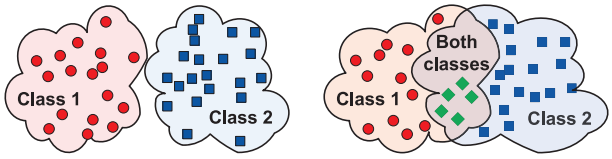
where $V_1$ is composed of the eigenvectors of $A$ corresponding to its positive eigenvalues. Note that, $A$ is semi-positive definite and thereby has no negative nonzero eigenvalues.

Because in many real applications such as computer vision and document content analysis, the number of features of a data set is often much larger than that of data points, *i.e.*, $p > n$, $S_w$ could also be rank deficient. In this case, we replace ${S_w}^{-\frac{1}{2}}$ in Eq. (1) by ${S_w}_+^{-\frac{1}{2}}$, which is computed in a same way as in Eqs. (10–11).

Finally, through an integral optimization objective, together with the supervision information contained in training data, data attributes and pairwise similarities on all the data points including those unlabeled are integrated by DLE. Therefore, DLE provides a framework for semi-supervised learning, which is summarized in Table 1.

## Multi-label DLE for Multi-Label Classification

*Multi-label classification* has attracted increasing attention in the past few years as it arises in a lot of real life applications such as image annotation, document categorization, *etc*. Different from traditional *single-label (multi-class) classification* where each object belongs to exact one class, multi-label classification deals with problems where each object may be associated with more than one class label, which makes it not trivial to directly apply state-of-the-art single-label classification algorithms to solve multi-label classification problems. On the other hand, since the multiple labels share the same input space, and the semantics conveyed by different labels are usually correlated, it is thereby beneficial to exploit the correlations among different labels to improve the overall classification accuracy. Therefore, we extend the proposed DLE to multi-label classification by first solving the ambiguity problem in computing the scatter matrices caused by data points with multiple labels and then incorporating the label correlations to achieve enhanced classification performance.




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

Figure 1: (a) A traditional single-label classification problem. Each data point distinctly belongs to only one class. (b) A typical multi-label classification problem. The data points denoted by green diamonds belong to more than one class, which cause the ambiguity in scatter matrices calculations.

### Class-wise Scatter Matrices for Multi-Label Classification

In traditional single-label multi-class classification, the scatter matrices $S_b$, $S_w$ and $S_t$ used in standard LDA are well defined as per the geometrical dispersion of data points. These definitions, however, do not apply to multi-label case, because a data point with multiple labels belong to different classes at the same time, how much it should contribute to the between-class and within-class scatters remains unclear. As illustrated in Figure 1, the data points with multiple labels, denoted by the green diamonds in Figure 1(b), cause the ambiguity in computing the scatter matrices.

Therefore, instead of computing the scatter matrices from data points perspective as in standard LDA, we propose to formulate them by class-wise, *i.e.*, $S_b = \sum_{k=1}^K S_b(k)$, $S_w = \sum_{k=1}^K S_w(k)$, and $S_t = \sum_{k=1}^K S_t(k)$. In this way, the structural variances of the training data are represented more perspicuous and the construction of the scatter matrices turns out easier. Especially, the ambiguity, how much a data point with multiple labels should contribute to the scatter matrices, is avoided. The *multi-label between-class scatter matrix* is defined as:

$$S_b = \sum_{k=1}^K S_b^{(k)},\ S_b^{(k)} = \left(\sum_{i=1}^l Y_{ik}\right)(\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^T, \quad (12)$$

the *multi-label within-class scatter matrix* $S_w$ is defined as:

$$S_w = \sum_{k=1}^K S_w^{(k)},\ S_w^{(k)} = \sum_{i=1}^l Y_{ik}\,(\mathbf{x}_i - \mathbf{m}_k)\,(\mathbf{x}_i - \mathbf{m}_k)^T, \quad (13)$$

and the *multi-label total scatter matrix* is defined as:

$$S_t = \sum_{k=1}^K S_t^{(k)},\ S_t^{(k)} = \sum_{i=1}^l Y_{ik}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T, \quad (14)$$

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

$$\mathbf{m}_k = \frac{\sum_{i=1}^l Y_{ik}\mathbf{x}_i}{\sum_{i=1}^l Y_{ik}}, \quad \mathbf{m} = \frac{\sum_{k=1}^K \sum_{i=1}^l Y_{ik}\mathbf{x}_i}{\sum_{k=1}^K \sum_{i=1}^l Y_{ik}}. \quad (15)$$

Note that, the multi-label global mean $\mathbf{m}$ defined in Eq. (15) is different from the global mean in single-label sense as in standard LDA. The latter is defined as $\frac{1}{l}\sum_{i=1}^l \mathbf{x}_i$.

Table 2: Description of 6 UCI data sets.

| Data sets | Sizes | Classes | Dimensions |
|---|---|---|---|
| soybean | 562 | 19 | 35 |
| housing | 506 | 3 | 13 |
| protein | 116 | 6 | 20 |
| wine | 178 | 3 | 13 |
| balance | 625 | 3 | 4 |
| iris | 150 | 3 | 4 |

Table 3: Classification accuracy (in percentage) with 10% of data labeled.

| Data sets | Approaches | | | | | | |
|---|---|---|---|---|---|---|---|
| | SVM | CF | HF | LPP | KL | Transducer | DLE |
| soybean | 70.3 | 36.2 | 52.3 | 51.3 | 76.5 | 75.2 | **88.4** |
| housing | 53.2 | 54.4 | 52.3 | 56.1 | 55.8 | 56.1 | **59.3** |
| protein | 32.7 | 27.6 | 27.4 | 31.5 | 38.2 | 38.5 | **41.7** |
| wine | 73.6 | 43.0 | 39.6 | 74.5 | 71.4 | 74.5 | **85.9** |
| balance | 47.2 | 47.5 | 46.6 | 44.5 | 47.1 | 46.8 | **52.1** |
| iris | 58.4 | 40.0 | 33.3 | 63.2 | 61.9 | 62.3 | **75.3** |

When applied to single–label classification, the multi–label scatter matrices, $S_b$, $S_w$, and $S_t$, defined in Eqs. (12–14), are reduced to their corresponding counterparts in standard LDA. Most importantly, in standard LDA, $S_t = S_b + S_w$, which is still held by the multi-label scatter matrices.

With the application of the class-wise scatter matrices defined in Eqs. (12–13) in the optimization objective of proposed DLE in Eq. (1), the ambiguity problem is solved, and in the meantime the main advantage of standard LDA, representing data with the maximized separability, is preserved.

### Label Correlation Enhanced Pairwise Similarity

Existing embedding algorithms construct the pairwise similarity matrix ($W$) only from data attributes ($X$), while the label correlations are overlooked, which, however, is indeed useful for class membership inference in multi-label classification. Thus, we propose a *label correlation enhanced pairwise similarity* construction scheme to make use of them as:

$$W = W_{\mathrm{X}} + \beta W_{\mathrm{L}}, \tag{16}$$

where $W_{\mathrm{X}}$ could be constructed from attribute data as in previous embedding algorithms or obtained directly from experimental observations, and $W_{\mathrm{L}}$ is the label similarity matrix. $\beta$ is a parameter determining how much the pairwise relationship should be biased by label similarities, which is empirically selected as $\beta = \frac{\sum_{i,j,i\neq j} W_{\mathrm{X}}(i,j)}{\sum_{i,j,i\neq j} W_{\mathrm{L}}(i,j)}$.

$W_{\mathrm{X}}$ is constructed using Gaussian function as in most of existing approaches:

$$W_{\mathrm{X}}(i,j) = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma\right), \tag{17}$$

where $\sigma$ is the length scale hyperparameter and fine tuned by 5-fold cross validation in the current work.

$W_{\mathrm{L}}$ encodes the label correlations via assessing the label similarities between data points. The simplest measure of label similarity is the label overlap of two data points computed by $\mathbf{y}_i^T \mathbf{y}_j$. The bigger the overlap is, and the more similar the two data points are. The problem with this straightforward similarity measurement is that it treats all classes independent and therefore is unable to explore the correlations among them. In particular, it will give zero similarity whenever two data points do not share any label. However, two data points with no common label can still be strongly related if their attached labels are highly correlated. Therefore, instead of computing label similarity by dot product, we compute it by $\mathbf{y}_i^T\, C\, \mathbf{y}_j$ with normalization:

$$W_{\mathrm{L}}(i,j) = \frac{\mathbf{y}_i^T\, C\, \mathbf{y}_j}{\|\mathbf{y}_i\|\,\|\mathbf{y}_j\|}, \tag{18}$$

where $C$ is a symmetric matrix and its entry $C_{kl}$ captures the correlation between the $k$th and $l$th classes. As the number of shared data points attached to two classes measures how closely they are related, we use the cosine similarity to quantify the label correlations. Let $Y = [\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(K)}]$, $\mathbf{y}_{(k)} \in \mathbb{R}^n (1 \leq k \leq K)$ is an $n$-vector, which is a class-wise label indicator vector for the $k$th class. Note that, $\mathbf{y}_{(k)}$ is different from the data-point-wise label indicator vectors $\mathbf{y}_i$ which is a $K$-vector. We define the label correlation matrix, $C \in \mathbb{R}^{K\times K}$, to characterize the label correlation between two classes as following:

$$C\,(k,l) = \cos(\mathbf{y}_{(k)}, \mathbf{y}_{(l)}) = \frac{\langle \mathbf{y}_{(k)}, \mathbf{y}_{(l)} \rangle}{\|\mathbf{y}_{(k)}\|\,\|\mathbf{y}_{(l)}\|}. \tag{19}$$

In order to compute $W_{\mathrm{L}}$ in Eq. (18), we first initialize the testing data points using 1NN method.

Apparently, $W_{\mathrm{L}}$ contains both supervision information and label correlations. Using the label correlation enhanced pairwise similarity $W$ defined in Eq. (16) to compute $A$ in Eq. (1), we incorporate the label correlations into the optimization objective of proposed DLE, and the overall multi-label classification performance is thereby boosted. Together with the utility of the class-wise scatter matrices $S_b$ and $S_w$ defined in Eqs. (12–13) into Eq. (1), we call the embedding computed from Eq. (1) as Multi-label DLE.

## Empirical Studies

### Semi-Supervised Learning Using DLE

We first evaluate the semi-supervised classification performance of proposed DLE by comparing to three state-of-the-art semi-supervised learning algorithms: (1) the consistent framework (CF) approach (Zhou et al. 2004) and (2) the harmonic function (HF) approach (Zhu, Ghahramani, and Lafferty 2003), (3) transducer approach (Joachims 2003), and two related embedding algorithms: LPP approach (He and Niyogi 2003) and embedding via K-means Laplacian clustering (KL) approach (Wang, Ding, and Li 2009). We also report the classification accuracy by Support Vector Machine (SVM) as a baseline. For the proposed DLE approach, LPP approach and KL approach, 1NN is used for classification after embedding. Six standard UCI data sets (Newman

Table 4: Classification accuracy with 10% correctly labeled data and another 5% incorrectly labeled data (noises).

| Data sets | Approaches | | | | |
|---|---|---|---|---|---|
| | CF | HF | LPP | KL | DLE |
| soybean | 36.2% | 54.8% | 50.1% | 71.2% | **85.1%** |
| housing | 54.4% | 49.9% | 53.4% | 53.9% | **57.3%** |
| protein | 27.6% | 23.8% | 30.9% | 37.4% | **39.1%** |
| wine | 43.0% | 40.1% | 74.4% | 71.9% | **83.6%** |
| balance | 47.2% | **64.4%** | 43.9% | 47.2% | 52.0% |
| iris | 38.0% | 50.0% | 60.1% | 57.1% | **71.9%** |

et al. 1998) are used in our evaluations, which are summarized in Table 2.

**Semi-supervised classification**. We randomly select 10% of data points, and treat them as labeled data and the rest as unlabeled data. We run the five compared approaches. To get good statistics, we rerun these trails 10 times such that the labeled data are different from each trial. Final results are the averages over these 10 trials and listed in Table 3. Over all six data sets, our DLE approach generally outperforms the other approaches, sometimes very significantly.

**Semi-supervised classification with noisy labels**. For each data set, 10% data points are randomly selected and given correct labels. Another 5% data points are randomly selected and are given incorrect labels, to emulate noises. All five learning methods are applied to the six data sets. Results of the averages of 10 trials of random samples are listed in Table 4. In general, accuracies in Table 4 are slightly worse than those in Table 3, as expected. Again, our DLE shows better performance than other approaches.

## Face Recognition Using DLE

In this subsection, we show the experimental results of applying our proposed DLE approach to face recognition on the following three benchmark data sets.

- **Yale** database contains 165 gray scale images of 15 individuals. There are 11 images per subject, one per different facial expression or configuration.
- **ORL** database Contains 10 different images of each of 40 distinct subjects. For some subjects, the images were taken at different times, varying the lighting, facial expressions and facial details.
- **PIE** database contains 41,368 images of 68 people, each person under 13 different poses, 43 different illumination conditions, and with 4 different expressions. In our experiments, we only use a subset containing 5 near frontal poses (C05, C07, C09, C27, C29) and all the images under different illuminations and expressions. We hence end up with 170 images for each individual.

Following standard computer vision experimental conventions, we resize all the face images to $32 \times 32$. Besides the proposed DLE approach, we also implement the following approaches:

- Nearest Neighbor Classifier (1NN): This method is implemented as the baseline method for comparison, where all the computations are performed in the original data space.

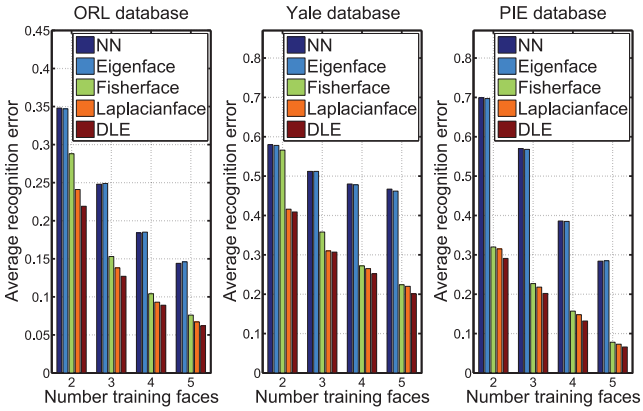



Figure 2: Face recognition results on 3 benchmark face data sets. The $x$-axis gives the number of images randomly selected from each class (person) for training. The $y$-axis gives the averaged recognition error.

- Eigenface: The face images are first projected by PCA and then the classification is implemented in the projected space with the 1NN classifier. The projected dimension is set using exhaustive search by 5-fold cross validation.
- Fisherface: The implementation is the same as in (Belhumeur et al. 1997).
- Laplacianface: The implementation is the same as in (He et al. 2005).

In our experiments, we first randomly select a certain number of face images from each subject. Those selected images will be used as training set and the remaining images will be used for testing. The recognition error averaged over 50 independent trials are shown in Figure 2, which demonstrate the effectiveness of our method.

## Multi-Label Classification Using Multi-label DLE

We use standard 5-fold cross validations to evaluate the multi-label classification performance of the proposed Multi-label DLE approach, and compare the experimental results with the following most recent multi-label classification methods: (1) Multi-Label Gaussian harmonic Function (MLGF) (Zha et al. 2008) method, (2) Semi-supervised learning by Sylvester Equation (SMSE) (Chen et al. 2008) method, (3) Multi-Label Least Square (MLLS) (Ji et al. 2008) method and (4) Multi-label Correlated Green's Function (MCGF) (Wang, Huang, and Ding 2009) method. For MLGF and SMSE methods, we follow the detailed algorithms as described in the original works. For MCGF and MLLS, we use the codes posted on the authors' websites.

We apply all approaches on the following three broadly used multi-label data sets from different domains.

- **TRECVID 2005** data set (Smeaton, Over, and Kraaij 2006) contains 137 broadcast news videos, which has 61901 sub-shots with 39 concepts (labels). We randomly sample the data set such that each concept has at least 100 video key frames. We extract 384-dimensional block-wise (64) color moments (mean and variance of each color band) from each image as features, *i.e.*, $\mathbf{x}_i \in \mathbb{R}^{384}$.

Table 5: Multi-label classification performance measured by "Average Precision" for the five compared approaches.

| Approaches | Data sets | | |
|---|---|---|---|
| | TRECVID | Yahoo | Music |
| MLGF | 10.8% | 14.2% | 23.5% |
| SMSE | 10.7% | 13.5% | 21.3% |
| MLLS | 23.7% | 27.6% | 31.2% |
| MCGF | 24.9% | 24.3% | 30.3% |
| Multi-label DLE | **28.1%** | **29.3%** | **37.7%** |

- **Music emotion** data set (Trohidis et al. 2008) comprises 593 songs with 6 emotions (labels).
- **Yahoo** data set is described in (Ueda and Saito 2002), which is a multi-topic web page data set compiled from 11 top-level topics in the "yahoo.com". Each web page is described as 37187-dimensional feature vector. We use the "science" topic as it has maximum number of labels, which contains 6345 web pages with 22 labels.

For performance evaluation, we adopt the widely-used metric, Average Precision (AP), as recommended by TRECVID (Smeaton, Over, and Kraaij 2006). We compute the precision for each class and average them over all the classes to obtain the AP to assess the overall performance.

Table 5 presents the overall classification performance comparisons of the five approaches by 5-fold cross validations on the three datassets. The results show that the proposed DLE approach clearly outperforms all the other methods, which justify the utility of class-wise scatter matrices in solving the ambiguity caused by data points with multiple labels and quantitatively demonstrate the usefulness of the incorporated label correlations in multi-label classification.

## Conclusions

In this work, we proposed a novel Discriminant Laplacian Embedding (DLE) approach to seek the shared structures between the *attributes* and *pairwise similarities* of a data set in a lower-dimensional subspace, in which the data separability is doubly reinforced by utilizing the information from the both data sources. With the integration of the supervision information from training data, the resulted embedding space is more discriminative. By solving the ambiguity problem in computing scatter matrices of standard LDA in multi-label scenarios and leveraging the label correlations, we extended DLE to multi-label classification with enhanced classification performance. In extensive experimental evaluations, promising results have been obtained, which demonstrate the effectiveness of our approaches.

## Acknowledgments

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

## References

Belhumeur, P.; Hespanha, J.; Kriegman, D.; et al. 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE TPAMI*.

Belkin, M., and Niyogi, P. 2002. Laplacian eigenmaps and spectral techniques for embedding and clustering. *NIPS'02*.

Chen, G.; Song, Y.; Wang, F.; and Zhang, C. 2008. Semi-supervised Multi-label Learning by Solving a Sylvester Equation. In *SDM'08*.

Chung, F. 1997. *Spectral graph theory*. Amer Mathematical Society.

Fukunaga, K. 1990. *Introduction to statistical pattern recognition*. Academic Press.

Hall, K. 1970. An r-dimensional quadratic placement algorithm. *Management Science* 219–229.

He, X., and Niyogi, P. 2003. Locality preserving projections. *NIPS'03*.

He, X.; Yan, S.; Hu, Y.; Niyogi, P.; and Zhang, H. 2005. Face recognition using laplacianfaces. *IEEE TPAMI*.

Ji, S.; Tang, L.; Yu, S.; and Ye, J. 2008. Extracting shared subspace for multi-label classification. In *SIGKDD'08*.

Joachims, T. 2003. Transductive learning via spectral graph partitioning. In *ICML*.

Jolliffe, I. 2002. *Principal component analysis*. Springer.

Newman, D.; Hettich, S.; Blake, C.; and Merz, C. 1998. UCI repository of machine learning databases.

Roweis, S., and Saul, L. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science*.

Smeaton, A. F.; Over, P.; and Kraaij, W. 2006. Evaluation campaigns and trecvid. In *MIR'06*.

Tenenbaum, J.; Silva, V.; and Langford, J. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science*.

Trohidis, K.; Tsoumakas, G.; Kalliris, G.; and Vlahavas, I. 2008. Multilabel classification of music into emotions. In *ISMIR'08*.

Ueda, N., and Saito, K. 2002. Single-shot detection of multiple categories of text using parametric mixture models. In *SIGKDD'02*.

Wang, F.; Ding, C.; and Li, T. 2009. Integrated KL (K-means-Laplacian) Clustering: A New Clustering Approach by Combining Attribute Data and Pairwise Relations. *SDM'09*.

Wang, H.; Huang, H.; and Ding, C. 2009. Image Annotation Using Multi-label Correlated Greens Function. In *ICCV'09*.

Zha, Z.; Mei, T.; Wang, J.; Wang, Z.; and Hua, X. 2008. Graph-based semi-supervised learning with multi-label. In *ICME'08*.

Zhou, D.; Bousquet, O.; Lal, T.; Weston, J.; and Schölkopf, B. 2004. Learning with local and global consistency. *NIPS'04*.

Zhu, X.; Ghahramani, Z.; and Lafferty, J. 2003. Semi-supervised learning using Gaussian fields and harmonic functions. *ICML'03*.