Integrating Global and Local Structures: A Least Squares Framework for Dimensionality Reduction

Jianhui Chen, Jieping Ye
Computer Science and Engineering Department
Arizona State University
{jianhui.chen, jieping.ye}@asu.edu

Qi Li
Computer Science Department
Western Kentucky University
qi.li@wku.edu

Abstract

Linear Discriminant Analysis (LDA) is a popular statistical approach for dimensionality reduction. LDA captures the global geometric structure of the data by simultaneously maximizing the between-class distance and minimizing the within-class distance. However, local geometric structure has recently been shown to be effective for dimensionality reduction. In this paper, a novel dimensionality reduction algorithm is proposed, which integrates both global and local structures. The main contributions of this paper include: (1) We present a least squares formulation for dimensionality reduction, which facilities the integration of global and local structures; (2) We design an efficient model selection scheme for the optimal integration, which balances the tradeoff between the global and local structures; and (3) We present a detailed theoretical analysis on the intrinsic relationship between the proposed framework and LDA. Our extensive experimental studies on benchmark data sets show that the proposed integration framework is competitive with traditional dimensionality reduction algorithms, which use global or local structure only.

1. Introduction

Dimensionality reduction has a wide range of applications in computer vision, statistical learning, and pattern recognition [7, 12]. A well-known approach for supervised dimensionality reduction is Linear Discriminant Analysis (LDA) [9, 10]. It has been used widely in various applications, including face recognition and microarray expression data analysis [1, 8]. Given a training data set, LDA captures its global geometric structure by maximizing the between-class distance and minimizing the within-class distance simultaneously, thus achieving maximum class discrimination.

Local geometric structure has recently received much attention in dimensionality reduction [2]. The local structure of a training data set may be captured by a Laplacian matrix [5], which is constructed from an adjacency graph of the data. (Details can be found in Section 3.) Laplacian Eigenmaps [2] and Locality Preserving Projection (LPP) [13] are nonlinear and linear dimensionality reduction algorithms, respectively, which compute a (locality preserving) low-dimensional manifold based on the graph Laplacian.

Structures of real-world data are often complex, and a single characterization (either global or local) may not be sufficient to represent the underlying true structures [6]. In this paper, we propose a novel dimensionality reduction algorithm, which integrates both global and local structures. First, we present a least squares formulation for LDA, under which the integration of global and local structures can be modeled as a regularized least squares problem. More specifically, LDA is regularized by a penalty term based on the graph Laplacian. A tuning parameter is employed to balance the tradeoff between global and local structures. We thus name the proposed algorithm *LapLDA*. It is worthwhile to note that the idea of regularization has a rich mathematical history going back to Tikhonov [17], where it is used for solving ill-posed inverse problems. Regularization is the key to many other machine learning methods such as Support Vector Machines (SVM) [18], spline fitting [19], etc. The use of the graph Laplacian as the regularization has been studied in [3] in the context of regression and SVM.

Second, we design an efficient model-selection algorithm to estimate the optimal tuning parameter involved in the integration framework. For high-dimensional data, the computational cost of solving a regularized least squares problem may be high. Our theoretical analysis shows that the regularization in the least squares formulation only takes place on a small size matrix. Based on this analysis, we develop an efficient algorithm for the estimation of the optimal tuning parameter from a given candidate set. Third, we present a detailed theoretical analysis on the intrinsic relationship between LapLDA and uncorrelated LDA [21]. (Uncorrelated LDA is an extension of classical LDA to deal with the singularity problem [15].) We show that under a

1-4244-1180-7/07/$25.00 ©2007 IEEE

mild condition, LapLDA based on a specific Laplacian matrix is equivalent to ULDA. This equivalence relationship provides new insights into the proposed integration framework.

We present extensive experimental studies to evaluate the effectiveness of LapLDA using six benchmark data sets, in comparison with ULDA and LPP, which use global or local structure only. The six data sets used in the evaluation cover a wide range of dimensions. They also reflect significant variations of the structure of data. For example, the use of local structure (in LPP) achieves better classification performance than the use of global structure (in ULDA) in some data sets, while it is not the case in other data sets. We observe, however, that the proposed integration framework outperforms both LPP and ULDA in all test date sets. This implies that global and local structures can be complementary to each other, even though one of them may be more important than the other one. These results validate the value of the integration of global and local structures in the proposed framework.

2. Linear Discriminant Analysis

Given a data set X consisting of n data points $\{x_1, x_2, \cdots, x_n\} \in \mathbb{R}^d$, classical LDA computes an optimal linear transformation $G \in \mathbb{R}^{d\times r}$ that maps the data onto a lower-dimensional subspace as follows:

$$x_j \in \mathbb{R}^d \rightarrow x_j^L = G^T x_j \in \mathbb{R}^r (r < d).$$

Let X be partitioned into k classes as $X = [X_1, X_2, \cdots, X_k]$, where $X_i \in \mathbb{R}^{d\times n_i}$, and n_i denotes the size of the i-th class. In discriminant analysis [10], three scatter matrices, i.e., *within-class*, *between-class*, and *total* scatter matrices are defined as follows:

$$S_w = \frac{1}{n}\sum_{i=1}^{k}\sum_{x\in C_i}(x - c^{(i)})(x - c^{(i)})^T, \quad (1)$$

$$S_b = \frac{1}{n}\sum_{i=1}^{k} n_i (c^{(i)} - c)(c^{(i)} - c)^T, \quad (2)$$

$$S_t = \frac{1}{n}\sum_{j=1}^{n}(x_j - c)(x_j - c)^T, \quad (3)$$

where $x \in C_i$ implies that x belongs to the i-th class, $c^{(i)}$ is the centroid of the i-th class, and c is the global centroid. It is easy to verify that $S_t = S_b + S_w$.

The scatter matrices S_w, S_b, and S_t can also be expressed as follows:

$$S_w = H_w H_w^T, \quad S_b = H_b H_b^T, \quad S_t = H_t H_t^T, \quad (4)$$

where matrices H_w, H_b, and H_t are given by

$$H_w = \frac{1}{\sqrt{n}}[X_1 - c^{(1)}e^{(1)^T}, \cdots, X_k - c^{(k)}e^{(k)^T}], \quad (5)$$

$$H_b = \frac{1}{\sqrt{n}}[\sqrt{n_1}(c^{(1)} - c), \cdots, \sqrt{n_k}(c^{(k)} - c)], \quad (6)$$

$$H_t = \frac{1}{\sqrt{n}}(X - ce^T), \quad (7)$$

and $e^{(i)}$ for all i and e are vectors of all ones with appropriate lengths. It follows that the trace of S_w measures the within-class cohesion, while the trace of S_b measures the between-class separation. In the lower-dimensional subspace resulting from the linear transformation G, the scatter matrices S_w, S_b, and S_t become $G^T S_w G$, $G^T S_b G$, and $G^T S_t G$, respectively.

An optimal transformation G^* would maximize the between-class distance and minimize the within-class distance simultaneously. In classical LDA, G^* is commonly computed by minimizing $\mathrm{tr}((G^T S_w G)^{-1} G^T S_b G)$, where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix [11]. This is equivalent to solving the following optimization problem:

$$G^* = \arg\max_G \left\{\mathrm{tr}((G^T S_t G)^{-1} G^T S_b G)\right\}, \quad (8)$$

due to the fact that $S_t = S_b + S_w$. The optimization problem in Eq. (8) is equivalent to finding all the eigenvectors x that satisfy $S_b x = \lambda S_t x$, for $\lambda \neq 0$ [10]. The optimal G^* can be obtained by applying an eigen-decomposition on the matrix $S_t^{-1} S_b$, provided that S_t is nonsingular. There are at most $k - 1$ eigenvectors corresponding to nonzero eigenvalues, since the rank of S_b is bounded from above by $k - 1$. Therefore, the reduced dimension of classical LDA is at most $k - 1$.

For high-dimensional data where the sample size may be smaller than the feature dimension, S_t can be singular and classical LDA fails. Many LDA extensions (see [21] for an overview) have been proposed in the past to overcome the singularity problem, including PCA+LDA, regularized LDA, Uncorrelated LDA (ULDA), null space LDA, etc.

3. Laplacian Linear Discriminant Analysis

In this section, we propose a novel least squares framework for dimensionality reduction, which integrates global and local geometric structures. The local structure is incorporated into the framework through the graph Laplacian, which can be considered as a regularization term in the least squares formulation. Interestingly, there is a close relationship between the proposed formulation and LDA as shown in Section 4 below. We thus name the algorithm LapLDA, which stands for Laplacian LDA.

3.1. Graph Laplacian

Given a data set $\{x_i\}_{i=1}^n$, a weighted graph can be constructed where each node in the graph corresponds to a data point in the data set. The weight S_{ij} between two nodes x_i and x_j is commonly defined as follows:

$$S_{ij} = \begin{cases} \exp(-\frac{\|x_i - x_j\|^2}{\sigma}) & x_i \in N_K(x_j) \text{ or } x_j \in N_K(x_i) \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where both K and $\sigma > 0$ are parameters to be specified, and $x_i \in N_K(x_j)$ implies that x_i is among the K nearest neighbors of x_j [2]. Let S be the similarity matrix whose (i, j)-th entry is S_{ij}. To learn an appropriate representation $\{z_i\}_{i=1}^n$ with preserved locality structure, it is common to minimize the following objective function [2]:

$$\sum_{i,j} \|z_i - z_j\|^2 S_{ij}. \tag{10}$$

Intuitively, $\|z_i - z_j\|$, the distance between two data points will be small if S_{ij} is large, i.e., x_i and x_j are close to each other in the original space. Thus the locality structure is preserved.

Define the Laplacian matrix L as $L = D - S$, where D is a diagonal matrix whose diagonal entries are the column sums of S, that is, $D_{ii} = \sum_{j=1}^n S_{ij}$. Note that L is symmetric and positive semidefinite. It is easy to verify that

$$\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \|z_i - z_j\|^2 S_{ij} = \text{tr}(ZLZ^T), \tag{11}$$

where $Z = [z_1, \cdots, z_n]$.

3.2. The Least Squares Framework

The proposed least squares formulation for dimensionality reduction incorporates the local structure information via a regularization term defined as in Eq. (11). Mathematically, LapLDA computes an optimal weight matrix W^*, which solves the following optimization problem:

$$W^* = \arg\min_W \{\|X^T W - Y\|_F^2 + \lambda \text{tr}(W^T X L X^T W)\} \tag{12}$$

where $||\cdot||_F$ denotes the Frobenius norm [11] of a matrix, $\lambda \geq 0$ is a tuning parameter, Y is a class indicator matrix defined as follows [22]:

$$Y(ij) = \begin{cases} \sqrt{\frac{n}{n_j}} - \sqrt{\frac{n_j}{n}}, & \text{if } y_i = j \\ -\sqrt{\frac{n_j}{n}}, & \text{otherwise} \end{cases} \tag{13}$$

y_i is the class label of x_i, n_j is the sample size of the j-th class, and n is the total sample size of the data. Here we assume that the data matrix X has been centered (of zero mean, i.e., $\sum_{i=1}^n x_i = 0$). The indicator matrix Y (of zero mean) is chosen so that $XY = nH_b$, which leads to the least squares formulation of LDA as shown in the next section. It is easy to verify that the optimal solution to LapLDA is given by

$$W^* = \left(\lambda X L X^T + X X^T\right)^+ n H_b, \tag{14}$$

where M^+ denotes the pseudo-inverse [11] of M.

3.3. Efficient Model Selection

The tuning parameter λ in Eq. (12) plays an important role in LapLDA, which is generally selected from a large set of candidates via cross-validation. However, for high-dimensional data, the sizes of XLX^T and XX^T are large, thus cross-validation is computationally prohibitive. In the following, we propose an efficient algorithm for the estimation of the optimal value of λ from a large search space, thus facilitating efficient model selection for LapLDA.

Assuming the data $X \in \mathbb{R}^{d \times n}$ has been centered, we have $H_t = \frac{1}{\sqrt{n}} X$. It follows that the solution to LapLDA in Eq. (14) can be expressed as follows:

$$\begin{aligned} W^* &= (n\lambda H_t L H_t^T + n S_t)^+ n H_b \\ &= \left(H_t(\lambda L + I_n) H_t^T\right)^+ H_b. \end{aligned} \tag{15}$$

Let $H_t = U \Sigma V^T$ be the SVD of H_t, where $H_t \in \mathbb{R}^{d \times n}$, $U \in \mathbb{R}^{d \times d}$, $V \in \mathbb{R}^{n \times n}$, and $\Sigma \in \mathbb{R}^{d \times n}$. Denote $t = \text{rank}(H_t)$. Let $U_1 \in \mathbb{R}^{d \times t}$ and $V_1 \in \mathbb{R}^{n \times t}$ consist of the first t columns of U and V, respectively. Let the square matrix $\Sigma_t \in \mathbb{R}^{t \times t}$ consist of the first t rows and the first t columns of Σ. We have

$$H_t = U \Sigma V^T = U_1 \Sigma_t V_1^T. \tag{16}$$

It follows from Eq. (15) that

$$\begin{aligned} W^* &= U \begin{pmatrix} \left(\Sigma_t V_1^T (\lambda L + I_n) V_1 \Sigma_t\right)^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^T H_b \\ &= U_1 \left(\Sigma_t V_1^T (\lambda L + I_n) V_1 \Sigma_t\right)^{-1} U_1^T H_b. \end{aligned}$$

For cases where the data dimension (d) is much larger than the sample size (n), the matrix

$$\Sigma_t V_1^T (\lambda L + I_n) V_1 \Sigma_t \in \mathbb{R}^{t \times t}$$

is much smaller than the matrix

$$H_t(\lambda L + I_n) H_t^T \in \mathbb{R}^{d \times d}$$

in the original formulation, thus dramatically reducing the computational cost. Note that for cases where the sample size is much larger than the data dimension, the original formulation is more efficient.

4. Relationship between LapLDA and ULDA

In this section, we show a close relationship between LapLDA and Uncorrelated LDA (ULDA) [21], which was recently proposed as an extension of classical LDA to deal with the singularity problem. One key property of ULDA is that the features in the transformed space are uncorrelated, thus ensuring minimum redundancy among the features in the reduced subspace. It was shown [21] that the transformation of ULDA consists of the first q eigenvectors of $S_t^+ S_b$, where $q = \text{rank}(S_b)$. We show in the following that LapLDA with a specific Laplacian matrix is equivalent to ULDA under a mild condition C1:

$$\text{rank}(S_t) = \text{rank}(S_b) + \text{rank}(S_w), \tag{17}$$

which has been shown to hold for many high-dimensional data sets [23].

4.1. An Important Property of Condition C1

Let $H_t = U\Sigma V^T = U_1\Sigma_t V_1^T$ be the SVD of H_t as defined in Eq. (16). Denote $B = \Sigma_t^{-1} U_1^T H_b \in \mathbb{R}^{t\times k}$. Let

$$B = P\tilde{\Sigma}Q^T \tag{18}$$

be the SVD of B, where P and Q are orthogonal. Define

$$H = U \begin{pmatrix} \Sigma_t^{-1}P & 0 \\ 0 & I_{d-t} \end{pmatrix}. \tag{19}$$

The following result is crucial for the equivalence relationship between LapLDA and ULDA [22]:

Theorem 4.1. *Let S_t, S_b, S_w, H_t, H_b, and H be defined as above. Let $t = rank(S_t)$ and $q = rank(S_b)$. Assume condition C1 in Eq. (17) holds. Then H in Eq. (19) diagonalizes the scatter matrices as follows:*

$$H^T S_t H = \begin{pmatrix} I_t & 0 \\ 0 & 0 \end{pmatrix}, \quad H^T S_b H = \begin{pmatrix} \Sigma_b & 0 \\ 0 & 0 \end{pmatrix}, \tag{20}$$

$$H^T S_w H = \begin{pmatrix} \Sigma_w & 0 \\ 0 & 0 \end{pmatrix}, \tag{21}$$

where

$$\Sigma_b = \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix}, \quad \Sigma_w = \begin{pmatrix} 0 & 0 \\ 0 & I_{t-q} \end{pmatrix}. \tag{22}$$

4.2. LapLDA based on a Specific Laplacian Matrix

We construct the following similarity matrix S using the class label information:

$$S_{ij} = \begin{cases} \frac{1}{n_\ell} & \text{if } x_i, x_j \text{ are from the } \ell\text{-th class} \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

where n_ℓ is the size of the ℓ-th class. It can be verified that

$$XLX^T = nS_w. \tag{24}$$

Therefore, the optimal solution to LapLDA in Eq. (14) can be expressed as

$$\begin{aligned} W^* &= \left(\lambda XLX^T + XX^T\right)^+ nH_b \\ &= (\lambda S_w + S_t)^+ H_b. \end{aligned} \tag{25}$$

Following the definition of H in Eq. (19), we have

$$\begin{aligned} & U^T S_w U \\ = & \begin{pmatrix} \Sigma_t^{-1}P & 0 \\ 0 & I_{d-t} \end{pmatrix}^{-T} H^T S_w H \begin{pmatrix} \Sigma_t^{-1}P & 0 \\ 0 & I_{d-t} \end{pmatrix}^{-1} \\ = & \begin{pmatrix} \Sigma_t P & 0 \\ 0 & I_{d-t} \end{pmatrix} \begin{pmatrix} \Sigma_w & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} P^T\Sigma_t & 0 \\ 0 & I_{d-t} \end{pmatrix} \\ = & \begin{pmatrix} \Sigma_t P\Sigma_w P^T\Sigma_t & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned} \tag{26}$$

We can thus express the optimal weight matrix in Eq. (25) as follows:

$$\begin{aligned} W^* &= (\lambda S_w + S_t)^+ H_b \\ &= \left\{ U \begin{pmatrix} \Sigma_t^2 & 0 \\ 0 & 0 \end{pmatrix} U^T + \lambda U (U^T S_w U) U^T \right\}^+ H_b \\ &= U_1 \left(\lambda \Sigma_t P \Sigma_w P^T \Sigma_t + \Sigma_t^2\right)^{-1} U_1^T H_b \\ &= U_1 \Sigma_t^{-1} P \left(\lambda \Sigma_w + I_t\right)^{-1} P^T \Sigma_t^{-1} U_1^T H_b \\ &= U_1 \Sigma_t^{-1} P \left(\lambda \Sigma_w + I_t\right)^{-1} P^T B \\ &= U_1 \Sigma_t^{-1} P (\lambda \Sigma_w + I_t)^{-1} \tilde{\Sigma} Q^T, \end{aligned} \tag{27}$$

where the second equality follows since U is orthogonal, the third equality follows since H_b lies in the null space of S_t, the fourth equality follows since P is orthogonal, and the last two equalities follow from the definition of B and Eq. (18).

4.3. Equivalence between LapLDA and ULDA

It can be verified from Theorem 4.1 that $\tilde{\Sigma}\tilde{\Sigma}^T = \Sigma_b$, that is,

$$\tilde{\Sigma} = \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix}.$$

It follows that

$$P\tilde{\Sigma} = [P_q, P_{t-q}]\tilde{\Sigma} = [P_q, 0], \tag{28}$$

where P_q consists of the first q columns of P. It follows from Theorem 4.1 that the first q diagonal entries of the matrix Σ_w are zero. Thus

$$(I_t + \lambda\Sigma_w)^{-1}\tilde{\Sigma} = \tilde{\Sigma}. \tag{29}$$

Combining Eqs. (27, 28,29), we have

$$\begin{aligned} W^* &= U_1\Sigma_t^{-1}P(I_t+\lambda\Sigma_w)^{-1}\tilde{\Sigma}Q^T \\ &= U_1\Sigma_t^{-1}P\tilde{\Sigma}Q^T \\ &= [U_1\Sigma_t^{-1}P_q, 0]Q^T. \end{aligned} \quad (30)$$

On the other hand, the optimal solution to ULDA consists of the top eigenvectors of $S_t^+S_b$. We can decompose $S_t^+S_b$ as follows:

$$\begin{aligned} & S_t^+S_b \\ =\ & U\begin{pmatrix} (\Sigma_t^2)^{-1} & 0 \\ 0 & 0 \end{pmatrix}U^TH_bH_b^T \\ =\ & U\begin{pmatrix} (\Sigma_t^2)^{-1}U_1^TH_bH_b^TU_1 & 0 \\ 0 & 0 \end{pmatrix}U^T \\ =\ & U\begin{pmatrix} \Sigma_t^{-1}BB^T\Sigma_t & 0 \\ 0 & 0 \end{pmatrix}U^T \\ =\ & U\begin{pmatrix} \Sigma_t^{-1}P & 0 \\ 0 & I \end{pmatrix}\begin{pmatrix} \Sigma_b & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} P^T\Sigma_t & 0 \\ 0 & 0 \end{pmatrix}U^T, \end{aligned}$$

where the last two equalities follows from $\Sigma_b = \tilde{\Sigma}\tilde{\Sigma}^T$ and $B = \Sigma_t^{-1}U_1^TH_b = P\tilde{\Sigma}Q^T$. Thus, the optimal transformation of ULDA is given by

$$G^* = U_1\Sigma_t^{-1}P_q, \quad (31)$$

since only the first q diagonal entries of Σ_b are nonzero.

The K-Nearest-Neighbor (K-NN) algorithm based on the Euclidean distance is commonly applied as the classifier in the dimensionality reduced space of LDA. If we apply W^* for dimensionality reduction before K-NN, the matrix W^* is invariant of any orthogonal transformation, since any orthogonal transformation preserves all pairwise distances. Thus W^* is essentially equivalent to $[U_1\Sigma_t^{-1}P_q, 0]$ or $U_1\Sigma_t^{-1}P_q$, as the removal of zero columns does not change the pairwise distance either. Thus LapLDA is equivalent to ULDA.

5. Experiments

In this section, we experimentally evaluate LapLDA in terms of classification accuracy on six benchmark data sets. 5-fold cross-validation is used in LapLDA for model selection. The Nearest-Neighbor (NN) algorithm is employed for classification. We set $K = 5$ and $\sigma = 1$ for the construction of the similarity matrix S defined in Eq. (9).

5.1. Data Set

We use various types of data sets (see Table 1 for more details) in the experiment, including text documents (20Newsgroups[1]), images (USPS [14]), and several other data sets from UCI Machine Learning Repository [16] such as Soybean, Waveform, Satimage, and Letter. The random partitions of the data into training and test sets with fixed sizes are used in the evaluation below.

[1]http://people.csail.mit.edu/jrennie/20Newsgroups/

Data Set	Training	Test	Sample size	Dim	Class
USPS	750	2250	3000	256	10
20Newsgroups	240	960	1200	8298	4
Waveform	300	900	1200	40	3
Satimage	300	3300	3600	36	6
Letter(a-m)	260	3640	3900	16	13
Soybean	150	412	562	35	15

Table 1. Statistics of the test data sets.

5.2. Efficiency

To evaluate the efficiency of the algorithm, we apply LapLDA to 20Newsgroups, which has the largest number of dimensions among all data sets used in the experiment. Table 2 shows the computational time (in seconds) of LapLDA for different sizes of the candidate set, ranging from 1 to 1024. It is clear that the cost of LapLDA grows slowly as the candidate size increases. Because of the improved efficiency of the proposed LapLDA algorithm, it is practical to select the optimal value of λ from such a large candidate set. In the following experiment, we choose the optimal value of the tuning parameter from a candidate set of size 1024.

5.3. Classification Performance

In the following comparative study, we compare LapLDA, which integrates both global and local structures, with ULDA and LPP, which use either global or local structure only. For each data set, we randomly partition the data into a training set and a test set (with a fixed training and test sizes). To give a better estimation of classification accuracy, the partition procedure is repeated 30 times and the resulting accuracies are averaged. Detailed description for the partition is presented in Table 1.

The classification accuracies of the 30 different partitions for all six data sets are presented in Figure 1. We also report the mean accuracy and standard deviation of the 30 different partitions for each data set in Table 3. It is clear from the presented results that LapLDA outperforms LPP and ULDA for all of the six data sets used in the experiment. This conforms to our motivation of integrating both global and local geometric structures in LapLDA to improve dimensionality reduction and classification.

We can observe from Figure 1 that ULDA outperforms LPP on USPS, 20Newsgroups, and Waveform, while LPP outperforms ULDA on Satimage and Letter. This implies that the relative importance of global and local structures depends on specific applications. For example, in USPS, 20Newsgroups, and Waveform, the global structure may play a more important role for the classification, while in Satimage and Letter, the local structure may contain more important information. However, in all these cases, LapLDA outperforms both ULDA and LPP. This implies that global and local structures can be complementary to each other, even though one of them may dominate the

Size	1	2	4	8	16	32	64	128	256	512	1024
Time	8.78	8.86	8.91	9.20	9.73	10.23	11.31	13.64	18.66	28.41	49.25

Table 2. Computational time (in seconds) of LapLDA for different sizes of the candidate set using 20Newsgroups.

Data Set	LapLDA		LPP		ULDA	
	mean	std	mean	std	mean	std
USPS	84.52	0.89	77.87	1.44	81.57	0.94
20Newsgroups	80.34	1.56	50.25	3.67	77.33	2.12
Waveform	80.35	1.78	65.89	1.87	72.50	2.25
Satimage	83.47	1.46	80.18	1.26	75.15	1.41
Letter (a-m)	80.24	1.18	79.20	1.26	75.26	1.49
Soybean	86.61	2.76	84.24	2.48	85.15	2.53

Table 3. Comparison of classification accuracies (in percentage) of LapLDA, LPP, and ULDA. The mean and standard deviation of 30 different partitions are reported.

other one in certain applications. Therefore, integrating both global and local structures in LapLDA may be beneficial. This explains why LapLDA may outperform both LPP and ULDA by a large margin, which is the case for USPS and Waveform.

It is interesting to note that all three algorithms achieve comparable performance on Soybean (see Figure 1). This implies that in certain cases such as Soybean, global and local structures may capture similar information and integrating both structures does not help.

6. Discussion

In this paper, we mainly focus on the case where all training examples are labeled. Semi-supervised learning, which occupies the middle ground between supervised learning (in which all training examples are labeled) and unsupervised learning (in which no labeled data are given), has received increased interest in recent years, particularly because of application domains in which unlabeled data are plentiful, such as images, text, and bioinformatics [4, 20, 24, 25]. The proposed least squares framework can be naturally extended to deal with unlabeled data, as graph Laplacian is defined on all data points (labeled and unlabeled) [3].

We have conducted a preliminary study to evaluate semi-supervised LapLDA using both labeled and unlabeled data on the USPS handwritten digits. The result is summarized in Figure 2, where the x-axis denotes the number of unlabeled data points used in the training set per class and the y-axis denotes the classification accuracy on a separate test set. We tested four different cases when different numbers of labeled data points per class were used. We can observe from Figure 2 that the use of unlabeled data does help to improve the performance of LapLDA.

7. Conclusions

We propose in this paper a novel least squares formulation of LDA, called LapLDA, for dimensionality reduction and classification. LapLDA integrates both global and local geometric structures, where the local structure is captured by the graph Laplacian. An efficient model selection algorithm is also presented to estimate the optimal integration model specified by a tuning parameter. A further theoretical analysis on the intrinsic relationship between LapLDA and ULDA is presented. We evaluate the proposed LapLDA algorithm on a collection of benchmark data sets in terms of classification accuracy. LapLDA is competitive with both ULDA and LPP in all cases. This validates the integration of both global and local structures in LapLDA.

The current work focuses on linear dimensionality reduction. This may be less effective when there is nonlinearity in the data. One of our future work is to extend LapLDA to deal with nonlinear data using kernels.

Acknowledgments

This research is sponsored by the Center for Evolutionary Functional Genomics of the Biodesign Institute at ASU and by the National Science Foundation Grant IIS-0612069.

References

[1] P. Belhumeour, J. Hespanha, and D. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997. 1

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps and sepctral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems*, volume 15, 2001. 1, 3

[3] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: a geometric framework for learning from examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006. 1, 6

[4] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. The MIT Press, Cambridge, MA, 2006. 6

[5] F. R. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997. 1

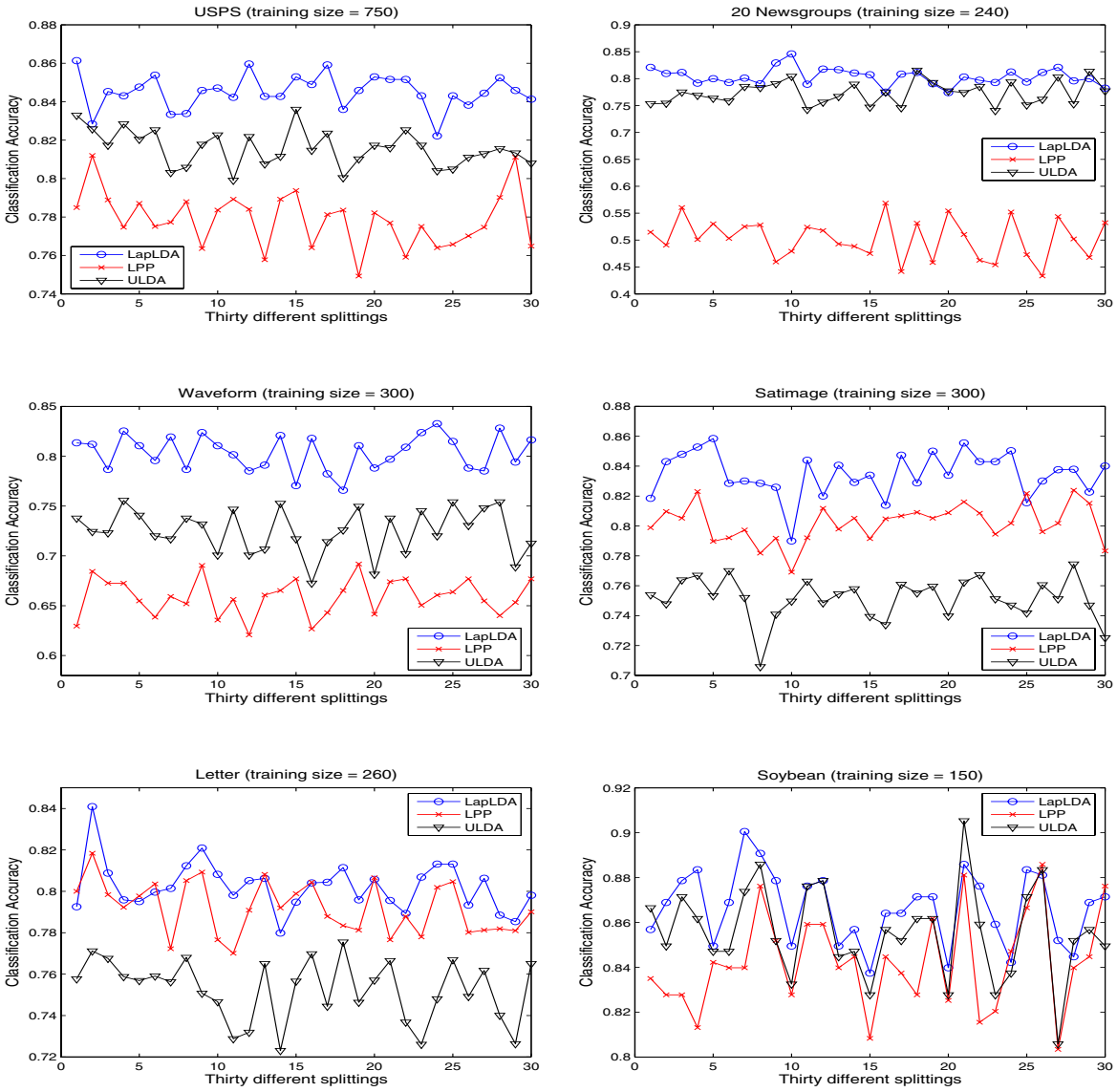

Figure 1. Comparison of LapLDA, LPP, and ULDA in classification accuracy using six benchmark data sets. The x-axis denotes 30 different partitions into training and test sets.

[6] V. de Silva and J. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In *Advances in Neural Information Processing Systems*, pages 705–712, 2002. 1

[7] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley Interscience, 2nd edition, 2000. 1

[8] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87, 2002. 1

[9] R. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936. 1

[10] K. Fukunaga. *Introduction to Statistical Pattern Classification*. Academic Press, San Diego, California, USA, 1990. 1, 2

[11] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, MD, USA, third edition, 1996. 2, 3

[12] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning : Data mining, Inference, and Prediction*. Springer, 2001. 1

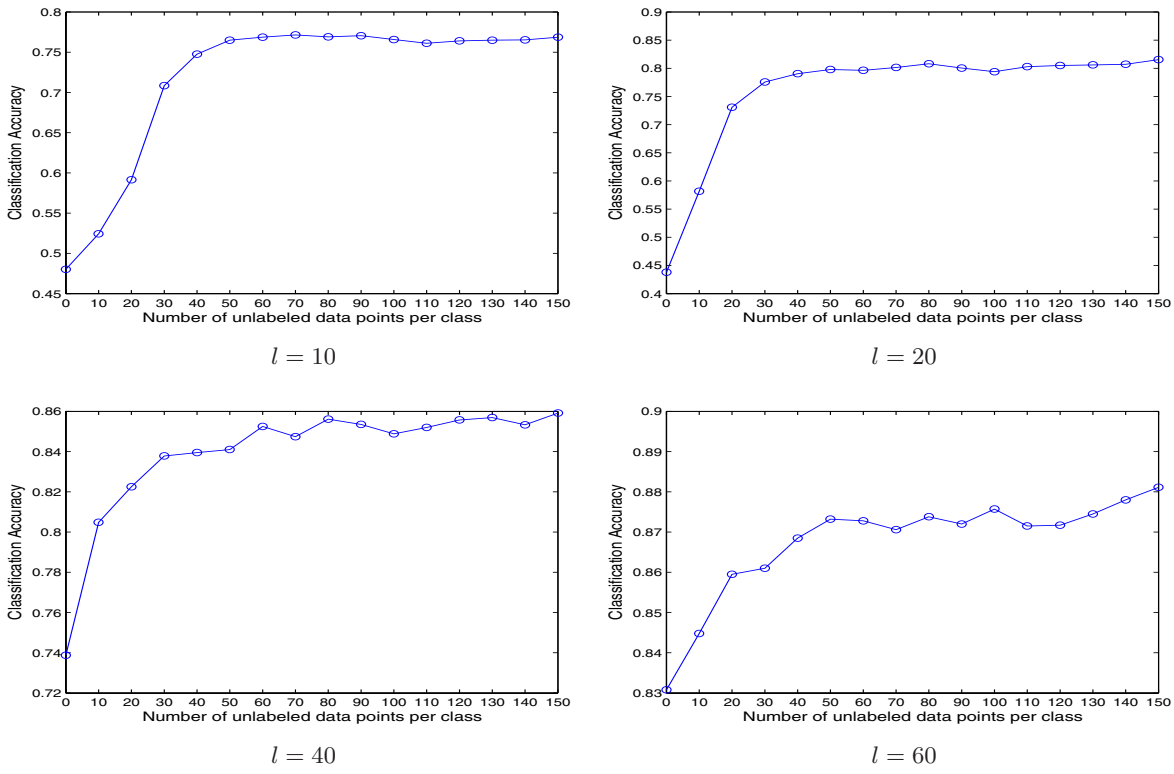

Figure 2. Effectiveness of semi-supervised LapLDA. The x-axis denotes the number of unlabeled data points per class used in the training set and the y-axis denotes the classification accuracy on a separate test set. l denotes the number of labeled data points per class.

[13] X. He and P. Niyogi. Locality preserving projection. In *Advances in Neural Information Processing Systems*, 2003. 1

[14] J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994. 5

[15] W. Krzanowski, P. Jonathan, W. McCarthy, and M. Thomas. Discriminant analysis with singular covariance matrices: methods and applications to spectroscopic data. *Applied Statistics*, 44:101–115, 1995. 1

[16] D. Newman, S. Hettich, C. Blake, and C. Merz. UCI repository of machine learning databases, 1998. 5

[17] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-posed problems*. John Wiley and Sons, Washington D.C., 1977. 1

[18] V. Vapnik. *Statistical learning theory*. Wiley, New York, 1998. 1

[19] G. Wahba. *Spline models for observational data*. Society for Industrial & Applied Mathematics, 1998. 1

[20] W. Weston, C. Leslie, E. Ie, D. Zhou, A. Elisseeff, and W. Noble. Semi-supervised protein classification using cluster kernels. *Bioinformatics*, 21(15):3241–3247, 2005. 6

[21] J. Ye. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *Journal of Machine Learning Research*, 6:483–502, 2005. 1, 2, 4

[22] J. Ye. Least squares linear discriminant analysis. In *Proceedings of the 24th International Conference on Machine Learning*, 2007. 3, 4

[23] J. Ye and T. Xiong. Computational and theoretical analysis of null space and orthogonal linear discriminant analysis. *Journal of Machine Learning Research*, 7:1183–1204, 2006. 4

[24] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, pages 321–328, 2003. 6

[25] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning*, pages 912–919, 2003. 6