DISCRIMINANT PAIRWISE LOCAL EMBEDDINGS

Konstantinos Bozas and Ebroul Izquierdo

School of EECS, Queen Mary University of London
{k.bozas, ebroul.izquierdo}@eecs.qmul.ac.uk

ABSTRACT

This paper introduces Discriminant Pairwise Local Embeddings (DPLE) a supervised dimensionality reduction technique that generates structure preserving discriminant subspaces. This objective is achieved through a convex optimization formulation where Euclidean distances between data pairs that belong to the same class are minimized, while those of pairs belonging to different classes are maximized. These pairwise relations are encoded in two matrices and weighted with the data affinity matrix to ensure local structure preservation. The discriminant efficiency of our technique is demonstrated in two popular applications, face and sketch recognition, where DPLE outperforms competitive manifold learning algorithms. A kernelized version of DPLE, that further enhances recognition accuracy, is also explained.

Index Terms— Dimensionality reduction, DPLE, manifold learning, sketch recognition, face recognition

1. INTRODUCTION

Dimensionality reduction or subspace learning is the transformation that maps data from a high-dimensional space into a meaningful low dimensional space. It has been widely used in recognition tasks to mitigate the inherent drawbacks of high-dimensional spaces. Real world data like images, videos and speech signals are by nature high-dimensional modalities. In order to efficiently process that data, we need to reduce its dimensionality. Furthermore, such real world data are accompanied by noise which affects the accuracy of classification algorithms. By exposing the intrinsic dimensionality of the input data, we can generate projection bases that are immune to noise. The benefits of dimensionality reduction include classification, visualization and compression of high-dimensional data [1].

One of the first and classic approaches to dimensionality reduction is the PCA algorithm that generates a subspace where data variance is maximized. PCA is an unsupervised technique, therefore does not produce discriminative subspaces. LDA [2] exploits the data labels and performs better in classification scenarios. PCA and LDA rely on assumptions on the data distributions which often do not hold for real world applications. LFDA [3] takes local structure of the data into account so multi-modal data can be embedded appropriately.

Manifold learning is the branch of dimensionality reduction that investigates the underlying manifold of data. Originated from ISOMAP [4], manifold learning techniques attempt to discover a low-dimensional manifold where the input data lie on. A famous example is the Swiss roll which is originally embedded in a three dimensional space, yet it easy to show by 'unfolding' it, that its points lie on a two dimensional manifold.

In the same spirit, LPP [5] and its variants [6, 7, 8, 9, 10] generate lower dimensional spaces that preserve the local neighborhood of the data, hence the restricting assumptions of PCA and LDA are avoided. LPP is an unsupervised technique, yet extensions have been published that make use of data labels. DLPP [9] incorporates in the optimization process the within and between scatter matrices to achieve class separability. ILPP [6], ARE [8] and max-margin MMP[10] are semi-supervised approaches obtaining label information from user feedback. ILPP updates its learned projection matrix according to user guidelines. MMP solves an eigenvalue problem that maximizes the margin between different labelled samples.

We present Discriminant Pairwise Local Embeddings (DPLE), a manifold learning algorithm inspired by LPP [5]. The main idea is to learn a discriminant subspace where the data will be better separated than in the original input space, without violating much its local neighbourhood. The latter ensures that the data will maintain their manifold structure in the learned subspace, so classification algorithms can generalize better. We form these goals in a convex optimization problem that can be efficiently solved through eigendecomposition. A kernelized version is also introduced to further enhance classification accuracy. Experiments on two datasets demonstrate the advantages of our technique.

DPLE’s objective is similar to that of LDE[7]/ARE, yet our formulation is different and the superiority of our technique is attributed to the following factors: a) LDE does not exploit the importance of influential samples, i.e. samples with many proximate neighbours guaranteed not to be outliers. DPLE utilizes this information in its objective function. b) ARE employs a non-flexible encoding scheme for the relationships between data pairs. It weights equally every pair
and does not take into account the distances of samples in the
original space. This approach fails to alleviate the influence
of noisy data pairs that belong to the same class but they are
far away in the feature space. DPLE handles this problem by
weighting these relationships with the affinity matrix.

2. DISCRIMINANT PAIRWISE LOCAL EMBEDDING

This section describes Discriminant Pairwise Local Embed-
dings (DPLE), a novel supervised dimensionality reduction
technique and its kernelized variant via the kernel trick [11].

2.1. Linear DPLE

Let \( n \) pairs of data samples and its associated labels
\((x_i, y_i), i = \{1, 2, \ldots, n\}\), where \( x_i \in \mathbb{R}^d \) represents a data
sample and \( y_i \in \{1, 2, \ldots, |C|\} \) is the label of the \( i \)-th sam-
ple. \( |C| \) is the total number of classes. Let \( X \in \mathbb{R}^{d \times n} \) be
the matrix of all samples. The \( i \)-th column of \( X \) is \( x_i \). Let
\( z_i \in \mathbb{R}^p (1 \leq p \leq d) \) be an embedded sample and \( p \) the
dimension of the embedding space. Since we investigate di-

dimensionality reduction scenarios, we usually require \( p \ll d \).

Linear dimensionality reduction is performed via the trans-
formation matrix \( W \in \mathbb{R}^{d \times p} \):

\[
z_i = W^T x_i
\]

(1)

The structure information of the data set is represented in
the affinity matrix \( A \). The matrix \( A \) captures similarities
different labels:

\[
A_{i,j} = \begin{cases} 
    e^{-||x_i-x_j||^2/2\sigma^2}, & \text{if } x_i \in \mathcal{N}_k(x_j) \text{ or } x_j \in \mathcal{N}_k(x_i) \\
    0, & \text{otherwise}
\end{cases}
\]

(2)

where \( \mathcal{N}_k(x) \) represents the set of \( k \)-nearest neighbours of \( x \).
A simpler alternative to (2) is to set \( A_{i,j} = 1 \) if \( x_i \) is a nearest
neighbor of \( x_j \) or vice versa; otherwise \( A_{i,j} = 0 \). In both
cases, a high value of \( A_{i,j} \) indicates that \( x_i \) and \( x_j \) lie close
in the defined metric space and a low value that they lie apart.

Based on the label information, we define two pairwise rela-
tion matrices. The same-label matrix \( A^{(s)} \) representing all
the sample pairs that share the same label and the different-
label matrix \( A^{(d)} \) representing all the sample pairs with dif-
ferent labels:

\[
A_{i,j}^{(s)} = \begin{cases} 
    A_{i,j}, & \text{if } y_i = y_j \\
    0, & \text{otherwise}
\end{cases}
\]

(3)

\[
A_{i,j}^{(d)} = \begin{cases} 
    A_{i,j}, & \text{if } y_i \neq y_j \\
    0, & \text{otherwise}
\end{cases}
\]

(4)

We observe from (3) and (4) that matrices \( A^{(s)} \) and \( A^{(d)} \)
are weighted with the affinity matrix \( A \). If we assign a constant
value to similar and dissimilar pairs as in [8]; for instance if
\( A_{i,j}^{(s)} = 1 \) when \( y_i = y_j \) and \( A_{i,j}^{(d)} = 1 \) when \( y_i \neq y_j \), then
all the sample pairs will have equal weights resulting in loss
of structure information. Instead, by employing the affinity
matrix we assign an ‘importance’ value to each pair. Samples
that lie close in the original input space are more significant
and are imposed to lie close in the embedding space. On the
other hand, pairs that are apart in the original space are either
ignored or slightly contribute to the optimal solution. This
idea is similar to the local variant of LDA [3], yet employed
in a different learning framework. We suggest the following optimization problem:

\[
\arg \max_W \frac{1}{2} \sum_{i,j} (W^T x_i - W^T x_j)^2 \left( A_{i,j}^{(d)} - \gamma A_{i,j}^{(s)} \right)
\]

subject to: \( W^T X D^{(s)} X^T W = I \)

(5)

where \( D_{i,i} = \sum_{j=1}^n A_{i,j} \) is a diagonal matrix consisted of
the row sums of \( A \) and \( \gamma \) is a scalar to compensate for any
imbalance of different number of pair samples between
\( A^{(d)} \) and \( A^{(s)} \).

The above formulation minimizes the Euclidean distances
between all sample pairs that belong to the same cate-
gory through matrix \( A^{(s)} \) and in the same time maximizes
those between pairs belonging to different classes through
matrix \( A^{(d)} \). We have previously seen that each pair
relation is weighted by the affinity matrix \( A \), therefore
the intrinsic structure of data is maintained. The constrain
\( W^T X D^{(s)} X^T W = I \) is imposed to avoid the trivial solution
\( W = 0 \) and each entry \( D_{i,i} \) provides a measure of importance
to the embedded sample \( z_i = W^T x_i \).

The objective function in (5) can be rewritten as follows
using linear algebra properties:

\[
\arg \max_W J(W) = W^T X \left( L^{(d)} - \gamma L^{(s)} \right) X^T W
\]

subject to: \( W^T X D^{(s)} X^T W = I \)

(6)

where \( L^{(s)} = D^{(s)} - A^{(s)} \) and \( L^{(d)} = D^{(d)} - A^{(d)} \) are the
Laplacian matrices of \( A^{(s)} \) and \( A^{(d)} \) respectively.

We apply the Lagrange multipliers to the above problem
and set the derivative with respect to \( W \) to zero:

\[
X \left[ L^{(d)} - L^{(s)} \right] X^T W = \lambda X D^{(s)} X^T W
\]

(7)

The result is a generalized eigenvalue problem and since
\( L^{(s)} \), \( L^{(d)} \) and \( D \) are symmetric semi-definite matrices all the
eigenvalues are real positive numbers.

The optimal projection matrix \( W_{DPLE} \) is given by:

\[
W_{DPLE} = \left[ \sqrt{\lambda_1} w_1 \mid \sqrt{\lambda_2} w_2 \mid \cdots \mid \sqrt{\lambda_p} w_p \right]
\]

(8)

where \( \{ w_i \}_{i=1}^p \) are the generalized eigenvectors associated
with the \( p \) largest eigenvalues \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \) of (7).
**Algorithm 1: DPLE embedding**

**Data:** \((x_i, y_i) \ i \in \{1, 2, \ldots, n\}, \gamma, p\)

**Result:** Projection matrix: \(W_{DPLE}\)

1. Compute affinity matrix \(A\) according to (2).
2. Compute matrices \(A^{(s)}\) and \(A^{(d)}\) from (3) and (4).
3. Solve the generalized eigenproblem of (7).
4. Form the columns of \(W_{DPLE}\) from the eigenvectors of (7) corresponding to the largest eigenvalues.

The steps of DPLE are summarized in Algorithm 1. DPLE exploits the labelled information encoded through the latter is ensured by the leverage of the affinity matrix bases without violating the intrinsic structure of the data. The latter is ensured by the leverage of the affinity matrix which weights accordingly each sample pair. The embedded data lie on a discriminative semantic manifold which preserves local geometric relations. As a result classes become better separated in the learned subspace.

### 2.2. Kernel DPLE

In most real world applications, data in the original input space cannot be linearly separated, due to it is being generated from non-linear processes. In such cases, linear algorithms like DPLE fail to produce efficient embedding spaces. We show that by using the kernel trick [11], we can generate a non-linear map from the original high-dimensional feature space to a lower-dimensional manifold where non-linear data can be efficiently represented.

Let \(\phi: \mathbb{R}^d \rightarrow \mathcal{H}\) be a non-linear map function, mapping the Euclidean space \(\mathbb{R}^d\) to Hilbert space \(\mathcal{H}\). In Hilbert space the eigenvector problem of (7) becomes:

\[
\phi(X) \begin{bmatrix} L^{(d)} - L^{(s)} \end{bmatrix} \phi(X)^\top w = \lambda \phi(X) D \phi(X)^\top w \quad (9)
\]

There is no easy way to directly compute the mapping \(\phi(X)\), yet we can employ inner products of mapped data to solve the problem. We define the inner products of the mapped data as:

\[
K(x_i, x_j) = \phi(x_i)^\top \phi(x_j) \quad (10)
\]

The eigenvectors of (9) are linear combinations of \(\phi(x_1), \phi(x_2), \ldots, \phi(x_n)\), hence we can write:

\[
w = \sum_{i=1}^{n} \alpha_i \phi(x_i) = \phi(X) \alpha \quad (11)
\]

where \(\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_n]^\top \in \mathbb{R}^n\). Using (11) it is easy to obtain the kernelized eigenvalue problem:

\[
K \begin{bmatrix} L^{(d)} - L^{(s)} \end{bmatrix} K \alpha = \lambda K D K \alpha \quad (12)
\]

As before, the optimal embedding is consisted from the \(p\) eigenvectors corresponding to largest eigenvalues of (12).

### 3. EXPERIMENTS AND RESULTS

In this section, the classification efficiency of DPLE is demonstrated. Our method is applied to two popular learning tasks, face recognition and sketch recognition, and compared against various well-known discriminant subspace learning algorithms.

#### 3.1. Datasets and experimental setup

The two datasets used in our evaluation are the ORL face database [12] and the sketch recognition database (SKETCH) of [13]. The ORL dataset includes 40 subjects with 10 grayscale images per subject. Following the preprocessing of [7], we resize each image to \(28 \times 28\) pixels and vectorize the outcome. We apply PCA to the image vectors and keep 98% of the information.

The SKETCH dataset of [13] encompasses 20,000 unique human-drawn sketches evenly distributed over 250 object categories. Each image depicts a binary sketch of a single object. All sketches are rescaled to a fixed size and centred in the image canvas to accommodate scale and translation invariance. The human accuracy on the above database is 73% which highlights the challenge for machine classification. We observe high inter-class and intra-class variability. Some classes are easily recognized while others regularly misclassified to categories with similar visual appearance. Moreover, an object can be sketched quite differently by various individuals a fact that contributes to aforementioned intra-class variations. Each sketch is represented by an ensemble of local features that capture the main gradient orientations of a local sketch region. The data are publicly available from the authors’ website and in this paper we use them as provided with no alternations.

We compare our method with the k-nn classifier in the original space denoted as (NN), the classic PCA and LDA algorithms and a collection of more sophisticated manifold learning techniques, namely LPP [5], LFDA [3] and LDE [7] along with kernelized versions for the last two. The recognition accuracy of the k-nn classifier in the learned subspace is reported. The parameters of each algorithm are empirically tuned for every dataset. In the kernelized version of the algorithms, we employ the rbf kernel with \(\sigma = 1\). In the ORL database we perform 5-fold cross-validation, whereas in SKETCH dataset we follow the protocol of [13] and perform 3-fold cross-validation with stratified sampling.

#### 3.2. Results

The evaluation results are illustrated in Table 1. NN accuracy indicates that AT&T dataset is easy. We observe that the discriminant manifold learning algorithms perform better than PCA, LDA and the unsupervised LPP. DPLE and KD- PLE achieves the highest recognition rates in this dataset.
Table 1. Best recognition rates in the evaluation datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>ORL</th>
<th>SKETCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>97.5%</td>
<td>45%</td>
</tr>
<tr>
<td>PCA</td>
<td>98% ($p = 32$)</td>
<td>41.97% ($p = 250$)</td>
</tr>
<tr>
<td>LDA</td>
<td>98% ($p = 24$)</td>
<td>41.2% ($p = 100$)</td>
</tr>
<tr>
<td>LPP</td>
<td>96.25% ($p = 20$)</td>
<td>41.74% ($p = 300$)</td>
</tr>
<tr>
<td>LFDA</td>
<td>98.5% ($p = 12$)</td>
<td>48.10% ($p = 120$)</td>
</tr>
<tr>
<td>LDE</td>
<td>98.5% ($p = 21$)</td>
<td>48.18% ($p = 120$)</td>
</tr>
<tr>
<td>KDPLE</td>
<td>99% ($p = 23$)</td>
<td>49.02% ($p = 100$)</td>
</tr>
<tr>
<td>KLFDA</td>
<td>99% ($p = 24$)</td>
<td>48.93% ($p = 95$)</td>
</tr>
<tr>
<td>KLDE</td>
<td>99.25% ($p = 21$)</td>
<td>52.64% ($p = 200$)</td>
</tr>
<tr>
<td>KDPLE</td>
<td>99.25% ($p = 23$)</td>
<td>53.70% ($p = 253$)</td>
</tr>
</tbody>
</table>

Fig. 1. Sketch recognition accuracy of kernelized algorithms across varying dimensionality using k-nn classification. KDPLE constantly outperforms the rest methods.

5. REFERENCES

[10] Xiaofei He, Deng Cai, and Jiawei Han, “Learning a maximum margin subspace for image retrieval,” Knowledge and Data Engineering, IEEE Transactions on, vol. 20, no. 2, pp. 189 –201, Feb. 2008.

4. CONCLUSIONS

We have presented DPLE, a supervised manifold learning algorithm that generates discriminant embeddings with a convex optimization process based on pairwise relations between the data. A non-linear variant of the algorithm is also illustrated. We have demonstrated the superiority of DPLE over competitive dimensionality reduction techniques in two recognition datasets. Future work could be concentrated on the online updating of the projection matrix upon new sample arrival.

Acknowledgments

This work was partially supported by EU project 3DLife under grant agreement FP7-247688.