

Fast Linear Discriminant Analysis Using Binary Bases

Feng Tang and Hai Tao

Department of Computer Engineering, University of California, Santa Cruz, USA
{tang|tao}@soe.ucsc.edu

Abstract

Linear Discriminant Analysis (LDA) is a widely used technique for pattern classification. It seeks the linear projection of the data to a low dimensional subspace where the data features can be modelled with maximal discriminative power. The main computation involved in LDA is the dot product between LDA base vector and the data which is costly element-wise floating point multiplications. In this paper, we present a fast linear discriminant analysis method called binary LDA, which possesses the desirable property that the subspace projection operation can be computed very efficiently. We investigate the LDA guided non-orthogonal binary subspace method to find the binary LDA bases, each of which is a linear combination of a small number of Haar-like box functions. The proposed approach is applied to face recognition. Experiments show that the discriminative power of binary LDA is preserved and the projection computation is significantly reduced.

1. Introduction

By finding the feature space that can best discriminate objects from others, discriminative methods have been widely used in pattern classification applications including face recognition [2], image retrieval [5], tracking [4]. Linear discriminant analysis (LDA) is a widely used discriminative method. It provides a linear projection of the data into a low dimensional subspace with the outcome of maximum between-class variance and minimum within-class variances. LDA has been used for face recognition which is commonly called “fisherface” [2].

The main computation in LDA is the dot product of a data vector with all the LDA base vectors. This can be computationally expensive especially when the original data dimension is high because it involves many floating point multiplications. Recently [7] has used Haar-like box functions as image features for face detection. In particular, [6] showed that an image can be represented as a linear combination of binary box functions to arbitrary preci-

sion. This representation is called non-orthogonal binary subspace (NBS) because the binary box base vectors are generally not orthogonal to each other. Examples of such binary box functions are shown in Figure 1. These binary box base vectors consist of one or two 2D rectangular areas. In these areas, the base image values are 1. The rest image area has value 0. The main benefit of using this form of base vectors is that the projection of the data can be computed very efficiently. The inner product of an input vector with these base vectors can be computed as the image intensity sum in one or two rectangular image areas. As shown in [7], only three or seven integer additions are needed by using the integral image. This has inspired us to find a subspace representation that has similar discriminative power as LDA but with highly reduced computation using binary box base vectors.



Figure 1. Three typical binary box functions. Left and middle are one-box functions and right is a symmetric two-box function.

Main contributions of this paper include:

- A novel efficient discriminative subspace method called binary LDA (B-LDA) which has comparable classification performance with LDA but with much reduced computation.
- An LDA guided NBS method to obtain the binary LDA bases each of which is spanned by binary box functions.
- The application of the binary LDA method to face recognition.

The rest of the paper is organized as follows: we introduce some background of LDA in section 2. In section 3, the B-LDA method for general object classification is presented. The application of the B-LDA in face recognition is demonstrated in section 4. We conclude the paper in section 5.

2. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) [2] is a class specific discriminative subspace representation that utilizes supervised learning to find a set of base vectors, denoted as \mathbf{w}_i , in such a way that the ratio of the between- and within-class scatters of the training sample set is maximized. This is equivalent to solving the following generalized eigenvalue problem.

$$\mathbf{W}_{opt} = \arg \max_{\mathbf{W}} \frac{|\mathbf{W}^T \mathbf{S}_b \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_w \mathbf{W}|} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M] \quad (1)$$

where $\{\mathbf{w}_i | 1 \leq i \leq M\}$ are the LDA subspace base vectors, M is the dimension of the subspace. \mathbf{S}_b and \mathbf{S}_w are the between- and within-class scatter matrices with the following forms,

$$\mathbf{S}_b = \sum_{i=1}^c N_i (\mu_i - \mu)(\mu_i - \mu)^T \quad (2)$$

$$\mathbf{S}_w = \sum_{i=1}^c \sum_{\mathbf{x}_k \in \mathbf{X}_i} (\mathbf{x}_k - \mu_i)(\mathbf{x}_k - \mu_i)^T \quad (3)$$

where c is the number of classes. $\mathbf{x} \in \mathbf{R}^N$ is a data sample. \mathbf{X}_i is the set of samples with class label i . μ_i is the mean for the all the samples with the class label i . N_i is the number of samples in the class i . When \mathbf{S}_w is non-singular, the base vectors \mathbf{W} sought in the above equation are the first M most “significant” eigenvectors of $\mathbf{S}_w^{-1} \mathbf{S}_b$ that correspond to the M largest eigenvalues $\{\lambda_i | 1 \leq i \leq M\}$. For a given test sample \mathbf{x} , we can obtain its representation in LDA subspace by a simple linear projection $\mathbf{W}^T \mathbf{x}$ because the LDA base vectors are orthogonal to each other. To avoid the so called small sample size (SSS) problem, the input \mathbf{x} is usually projected to a low dimensional PCA subspace and use the projection coefficients as input to LDA.

3. The approach

3.1. The problem formulation

In the original LDA, the problem is formulated to find the linear subspace that can best discriminate the data within class from other classes. While in the proposed binary LDA, we aim at finding a subspace representation that can preserve the discrimination power of the traditional LDA and at the same time, reduce the computation cost involved in the floating point dot product. The binary LDA is formulated as follows:

$$\mathbf{W}_{opt} = \arg \max_{\mathbf{W}} \frac{|\mathbf{W}^T \mathbf{S}_b \mathbf{W}|}{|\mathbf{W}^T \mathbf{S}_w \mathbf{W}|} - \beta \sum_i c(\mathbf{w}_i) \quad (4)$$

$$\text{subject to: } \mathbf{w}_i = \sum_j \alpha_j \phi_j$$

where ϕ_i is a binary box function. α_i is the coefficient. $c(\mathbf{w}_i)$ is the computational cost of the base vector \mathbf{w}_i . Basically, the objective function consists of two terms: the first term is the discriminative power term which is the ratio of the between- and within-class scatters; the second term is the computation cost term, with β as a positive weight to control the relative importance of the two terms. Since the B-LDA base vectors are represented as linear combination of a small number of box functions, there is no guarantee that they are orthogonal to each other, so the B-LDA subspace is a non-orthogonal subspace. The relation between LDA and B-LDA subspaces is illustrated in Figure 2.

3.2. The solution: LDA guided NBS

The search space for the optimization problem in Eq. 4 is huge because the solution can be any base vector that is a linear combination of any box functions from the binary dictionary D . Even for a small image of size 24×24 as used in our experiments, there are 134998 box functions in D . This makes it difficult to find the global optimal solution. We propose an LDA guided NBS method to find a sub-optimal solution efficiently by approximating iteratively computed LDA bases using binary box functions. By controlling the number of binary box functions used to approximate LDA bases which is closely related with the $c(\mathbf{w}_i)$ in Eq.4, we can obtain a good approximate solution.

Optimized orthogonal matching pursuit (OOMP) used in NBS [6] to find the NBS base vectors is a technique for computing adaptive signal expansion by iterative selection of base vectors from a dictionary. Such a dictionary $D = \{\phi_i\}_{i \in I}$ is usually non-orthogonal (binary box functions in our paper). The OOMP algorithm iteratively selects base vectors $\Phi_\Lambda = [\phi_{l_1}, \dots, \phi_{l_{|\Lambda|}}]$ according to the following procedure: Suppose that at iteration k the already selected k base vectors are defined by the index set $\Lambda_k = (l_i)_{i=1}^k$. To find the next base vector in iteration $k+1$, the OOMP prescribes to select the index l_{k+1} that minimizes the new approximation error:

$$\varepsilon_{k+1} = \min_i \frac{|\langle \gamma_i, \varepsilon_k \rangle|}{\|\gamma_i\|}, \quad \|\gamma_i\| \neq 0, \quad i \in \bar{\Lambda}_k \quad (5)$$

where $\varepsilon_k = \mathbf{x} - R_{\Phi_{\Lambda_k}}(\mathbf{x})$ is the approximation error using Φ_{Λ_k} and $\gamma_i = \phi_i - R_{\Phi_{\Lambda_k}}(\phi_i)$. $R_{\Phi_\Lambda}(\mathbf{x}) = \Phi_\Lambda (\Phi_\Lambda^T \Phi_\Lambda)^{-1} \Phi_\Lambda^T \mathbf{x}$ is the reconstruction of the signal \mathbf{x} using the non-orthogonal base vectors Λ_k . $\bar{\Lambda}_k$ is the subset of indices that are not selected in the previous iteration k , i.e. $\bar{\Lambda}_k = I - \Lambda_k$. An effective implementation of this optimization can be achieved by the forward adaptive bi-orthogonalization [1]. In essence, OOMP is a greedy algo-

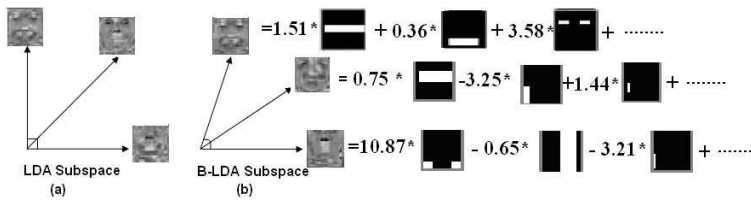


Figure 2. Relation of LDA subspace (orthogonal) and B-LDA subspace (non-orthogonal).

rithm that finds a sub-optimal decomposition of data vector using minimum number of base vectors in D .

In the LDA guided NBS, we denote the selected B-LDA base vectors up to iteration k as $\mathbf{W}_k = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k]$. This set is empty at the beginning. We start from the original LDA procedure to obtain the first principal component that captures the majority of the data variance. We call the first principal component the *Pre-LDA* vector, denoted as \mathbf{w}_1^- . NBS is then applied to approximate this vector as $\mathbf{w}_1 = \sum_{j=0}^{N_1} c_{j,1} \phi_{j,1}$. Then, in iteration k , the data \mathbf{X} is projected to the subspace spanned by the already selected B-LDA bases \mathbf{W}_{k-1} , and LDA is applied on the residual of the data $\mathbf{X} - R_{\mathbf{W}_{k-1}}(\mathbf{X})$ to obtain the next *Pre-LDA* \mathbf{w}_k^- which is again approximated using NBS. The approximation of *Pre-LDA* at iteration k is called the k -th *B-LDA base vector*. This procedure iterates until the desired number of B-LDA bases have been obtained or an error threshold is reached.

Generally, it takes a large number of box functions to represent each *Pre-BLDA* perfectly. However, the computational cost term in the objective function prefers a solution with fewer box functions. To make the optimization simpler, we enforce a computational cost constraint by finding the minimum number of box functions that satisfy:

$$(1 - \tau) \|\mathbf{w}\|^2 \leq \|\overline{\mathbf{w}}\|^2 \leq \|\mathbf{w}\|^2 \quad (6)$$

where $\overline{\mathbf{w}}$ is the reconstruction of \mathbf{w}_1^- using binary box functions. $\tau \in [0, 1]$ is the approximation error threshold that controls the precision. A smaller value of τ tends to produce a more accurate approximation. N is the dimension of the base vector. The comparison of LDA, pre-BLDA and B-PCA base vectors are shown in Figure 3. Figure 4 demonstrates the selected box functions used to approximate the first iteratively selected LDA base vectors.

Since these bases are linear combination of binary box functions, there is no guarantee that they are orthogonal to each other. As a result, the reconstruction process becomes $P_{\mathbf{W}}(\mathbf{x}) = \mathbf{W}(\mathbf{W}^T \mathbf{W})^{-1} \mathbf{W}^T \mathbf{x}$. This pseudo-inverse projection can be approximated using direct dot product (DNP): $P_{\mathbf{W}}(\mathbf{x}) = \mathbf{W}^T \mathbf{x}$. Experiments show that this approximation does not cause much performance deduction. It can be proved that the error between the direct dot product signal representation in B-LDA subspace and

that in LDA subspace has an upper bound.

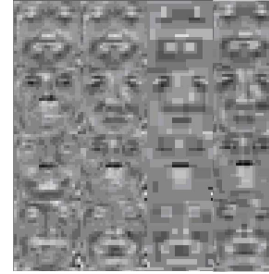


Figure 3. Comparison of the original LDA bases, pre-LDA bases, binary LDA bases ($\tau = 0.8$), binary LDA bases ($\tau = 0.5$), from left to right.

4. Experiments

We tested the proposed B-LDA method for face classification. B-LDA is applied on the training data to find the bases, then the testing image are projected onto these bases to obtain the feature vector, the classification is achieved using nearest neighbor.

4.1. Face recognition

For face recognition, 500 frontal view images from the FERET database were used. These images were spatially aligned and scaled to 24×24 pixels. The first 4 of these B-LDA bases are shown in Figure 3. Figure 4 shows the features used to approximate the first pre-LDA base vector. Figure 5 shows the classification performance using pseudo-inverse projection. As can be observed, in general, the performance increases with the number of base vectors used. Even with the approximation error τ to be 0.8 (very coarse approximation), the classification performance of B-LDA is comparable to LDA. Figure 6 is the comparison of the classification performance using direct dot product (DNP) approximation and the projection with pseudo-inverse. As can be observed, the smaller the τ , the smaller difference between DNP and pseudo-inverse projection.



Figure 4. Some of LDA guided OOMP selected box functions to approximate the first binary LDA base vector.

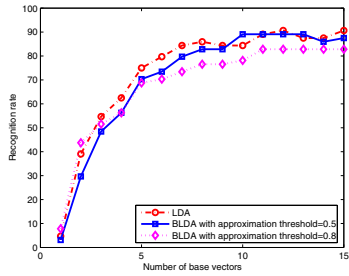


Figure 5. Comparison of the LDA and B-LDA performance with $\tau = 0.8$.

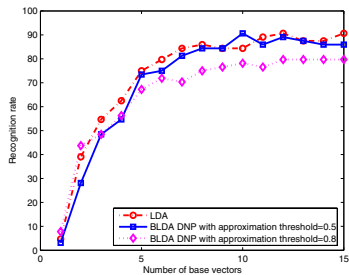


Figure 6. The recognition rate of the dot product and pseudo inverse with $\tau = 0.8$.

4.2. Speed Improvement

Suppose the data vector is of dimension N , and there are n LDA base vectors, the projection of a data vector to the LDA subspace takes $n \times N$ floating point multiplications and $n \times (N - 1) + (n - 1)$ floating point additions. But for B-LDA, each base vector is represented as a linear combination of binary box functions, i.e., $\mathbf{w}_i = \sum_j \alpha_j \phi_j$, the projection operation becomes, $\mathbf{w}_i^T \mathbf{x} = \sum_j \alpha_j (\phi_j^T \mathbf{x})$. The dot product between the image and box functions $\phi_j^T \mathbf{x}$ can be computed using 3 or 7 integer additions using integral image trick. So the computation for B-LDA is $n \times |\Lambda|$ floating point multiplications and between $3 \times n \times |\Lambda| + n - 1$ and $7 \times n \times |\Lambda| + n - 1$ additions, where $|\Lambda|$ is the number of binary box functions used to represent each LDA base vector. Since $|\Lambda| \ll N$, we have $n \times |\Lambda| \ll n \times N$, the computation of B-LDA is much less than LDA. Using B-LDA, the computation is reduced from $O(N)$ to constant,

which is only related to the number of box functions used to approximate each LDA base vectors. The experiment for speed improvement is carried out on a Pentium 4, 3.2GHz, 1G RAM machine, using C++ code. Fifteen base vectors are computed, for both LDA and B-LDA, and the time to project images onto each subspace is observed. The B-LDA used direct dot product projection. We tested 1000 samples and use compute the time of a single projection operation as the average and the results are listed in Table 1.

Table 1. Comparison of the computational cost between LDA and B-LDA projection operation.

Threshold: τ	0.5	0.8
#box functions	424	99
T_{LDA} (sec)	3.15×10^{-4}	
T_{ii} (integral image) (sec)	4.72×10^{-5}	
T_{BLDA} (sec)	8.13×10^{-6}	2.10×10^{-6}
Speedup($\frac{T_{LDA}}{T_{BLDA} + T_{ii}/N_{bases}}$)	27.97	60.23

5. Conclusion

A novel efficient discriminative method called B-LDA is presented in this paper. It inherits the properties of LDA in terms of discriminating data from different class while take advantages of the computational efficiency of non-orthogonal binary bases. We applied the B-LDA method to the face recognition. Promising results are demonstrated.

References

- [1] Andrieu, M. and Rebollo-Neira, L. "A swapping-based refinement of orthogonal matching pursuit strategies", *Signal Processing*, Vol (86,3), page 480-495 2006.
- [2] Belhumeur, P.N., Hespanha, J. and Kriegeman, D. "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE PAMI*, 1997.
- [3] Gilbert, A.C., Muthukrishnan, S., Strauss, M. "Approximation of functions over redundant dictionaries using coherence" in *Proceedings of the fourteenth annual ACM-SIAM symposium on discrete algorithms*, 2003.
- [4] Lin, R., Yang, M.H. and Levinson, S.E. "Object Tracking Using Incremental Fisher Discriminant Analysis" *ICPR* 2004.
- [5] Swets, D.L. and Weng, J. "Hierarchical discriminant analysis for image retrieval" *IEEE PAMI*. May 1999
- [6] Tao, H., Crabb, R. and Tang, F. "Non-orthogonal binary subspace and its applications in computer vision", in *ICCV* 2005.
- [7] Viola, P. and Jones, M. "Rapid object detection using a boosted cascade of simple features" in *CVPR* 2001