Linear Discriminant Dimensionality Reduction

Quanquan Gu   Zhenhui Li   Jiawei Han

Department of Computer Science
University of Illinois at Urbana-Champaign

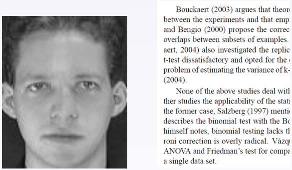
Sep 7, 2011

Outline

1. Background
2. Motivation
3. The Proposed Method
4. Experiments
5. Summary

Dimensionality Reduction

- In many real world application, data sample is represented by a high dimensional vector, e.g. face recognition, text classification
- Curse of dimensionality



(a) (b)

Figure: (a) face image: 92*112= 10304 pixels (b) text: about 20000 words in the vocabulary

- Dimensionality reduction: subspace learning, feature selection

Subspace Learning

- Transform the original input features to a lower dimensional subspace, but use all the original features
- e.g., Principal Component Analysis, Linear Discriminant Analysis (Belhumeur et al. PAMI'97), Locality Preserving Projection (He and Niyogi, NIPS'03)

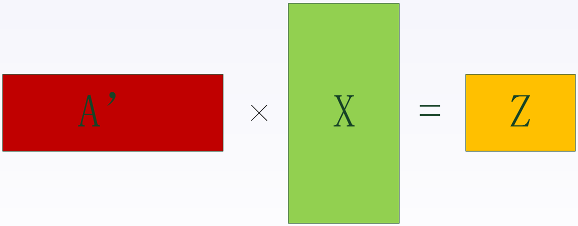


Figure: X is the original data matrix, A is the linear transformation matrix, Z is the projected data matrix in the subspace

Feature Selection

- Select a subset of most informative features
- e.g., Fisher Score (Duda and Stork ’01), Mutual Information, Information Gain (Guyon and Elisseeff, JMLR’03)

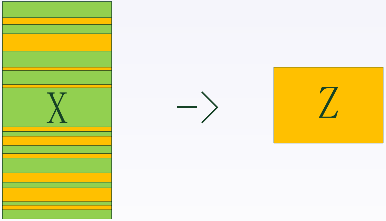


Figure: X is the original data matrix, the yellow rows in X are those selected features, Z is the data matrix with selected features

Outline

1 Background

2 Motivation

3 The Proposed Method

4 Experiments

5 Summary

Fisher Criterion

Fisher Criterion

Finding a feature representation by which the within-class distance is minimized and the between-class distance is maximized

- Fisher criterion plays an important role in dimensionality reduction.
- Based on Fisher criterion, two representative methods have been proposed.
  - *Linear Discriminant Analysis* (LDA), which is a subspace learning method.
  - *Fisher Score*, which is a feature selection method.

Linear Discriminant Analysis I

- Find a linear transformation $\mathbf{W} \in \mathbb{R}^{d \times m}$ that maps $\mathbf{x}_i$ in the d-dimensional space to a m-dimensional space, in which the between class scatter is maximized while the within-class scatter is minimized, i.e.,

$$\arg\max_{\mathbf{W}} \mathrm{tr}((\mathbf{W}^T \mathbf{S}_t \mathbf{W})^{-1} (\mathbf{W}^T \mathbf{S}_b \mathbf{W})), \tag{1}$$

- $\mathbf{S}_b$ and $\mathbf{S}_t$ are the between-class scatter matrix and total scatter matrix respectively, which are defined as

$$\begin{aligned} \mathbf{S}_b &= \sum_{k=1}^{c} n_k (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^T \\ \mathbf{S}_t &= \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T. \end{aligned} \tag{2}$$

Linear Discriminant Analysis II

- Advantage: Admit feature combination
- Disadvantage:
  - It transforms all the original features rather than only those useful ones
  - The resulting transformation is often difficult to interpret.

Fisher Score I

- Find a subset of features, such that in the data space spanned by the selected features, the distances between data points in different classes are as large as possible, while the distances between data points in the same class are as small as possible, i.e.,

$$\arg\max_{\mathbf{p}} \quad \mathrm{tr}\{(\mathrm{diag}(\mathbf{p})\mathbf{S}_t\mathrm{diag}(\mathbf{p}))^{-1}(\mathrm{diag}(\mathbf{p})\mathbf{S}_b\mathrm{diag}(\mathbf{p}))\},$$
$$\text{s.t.} \quad \mathbf{p} \in \{0,1\}^d, \mathbf{p}^T\mathbf{1} = m, \tag{3}$$

- $\mathbf{p}$ is an indicator variable, where $\mathbf{p} = (p_1, \ldots, p_d)^T$ and $p_i \in \{0,1\}, i = 1, \ldots, d$, to represent whether a feature is selected or not. $\mathrm{diag}(\mathbf{p})$ is a diagonal matrix whose diagonal elements are p_i's

Fisher Score II

- Advantage:
  - Able to find useful features
  - Interpretable
- Disadvantage: Does not admit feature combination like LDA does.
- LDA suffers from the problem which Fisher score does not have, while Fisher score has the limitation which LDA does not have.

Our Goal

- Integrate Fisher score and LDA in a unified framework
- Perform feature selection and subspace learning simultaneously based on Fisher criterion
- Inherit the advantages of Fisher score and LDA to overcome their individual disadvantages
- Be able to discard the irrelevant features and transform the relevant ones simultaneously

Outline

1 Background

2 Motivation

3 The Proposed Method

4 Experiments

5 Summary

Linear Discriminant Dimensionality Reduction

- Find a subset of features, based on which the learnt linear transformation via LDA maximizes the Fisher criterion.

$$\begin{aligned} \arg\max_{\mathbf{W},\mathbf{p}} \quad & \mathrm{tr}\{(\mathbf{W}^T \mathrm{diag}(\mathbf{p})\mathbf{S}_t \mathrm{diag}(\mathbf{p})\mathbf{W})^{-1}(\mathbf{W}^T \mathrm{diag}(\mathbf{p})\mathbf{S}_b \mathrm{diag}(\mathbf{p})\mathbf{W})\}, \\ \text{s.t.} \quad & \mathbf{p} \in \{0,1\}^d, \mathbf{p}^T\mathbf{1} = m, \end{aligned} \tag{4}$$

- Both Fisher score and LDA can be seen as the special cases of LDDR
  - $\mathbf{p} = \mathbf{1}$, Eq. (4) reduces to LDA
  - $\mathbf{W} = \mathbf{I}$, Eq. (4) degenerates to Fisher score
- The objective functions corresponding to LDA and Fisher score are lower bounds of the objective function of LDDR.
- It is a mixed integer programming, which is difficult to solve

Equivalent Formulation

Theorem

The optimal $\mathbf{p}$ *that maximizes the problem in Eq. (4) is the same as the optimal* $\mathbf{p}$ *that minimizes the following problem*

$$\begin{aligned} \arg\min_{\mathbf{p},\mathbf{W}} \quad & \frac{1}{2}||\mathbf{X}^T diag(\mathbf{p})\mathbf{W} - \mathbf{H}||_F^2 \\ s.t. \quad & \mathbf{p} \in \{0,1\}^d, \mathbf{p}^T\mathbf{1} = m, \end{aligned} \tag{5}$$

where $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_c] \in \mathbb{R}^{n\times c}$, *and* $\mathbf{h}_k$ *is a column vector whose* i*-th entry is given by*

$$h_{ik} = \begin{cases} \sqrt{\frac{n}{n_k}} - \sqrt{\frac{n_k}{n}}, & if\ \mathbf{y}_i = k \\ -\sqrt{\frac{n_k}{n}}, & otherwise. \end{cases} \tag{6}$$

In addition, the optimal $\mathbf{W}_1$ *of Eq. (4) and the optimal* $\mathbf{W}_2$ *of Eq. (5) have the following relation*

$$\mathbf{W}_2 = [\mathbf{W}_1, \mathbf{0}]\mathbf{Q}^T, \tag{7}$$

under the condition that $rank(\mathbf{S}_t) = rank(\mathbf{S}_b) + rank(\mathbf{S}_w)$ *and* $\mathbf{Q}$ *is an orthogonal matrix.*

Reformulation

- Suppose we find the optimal solution of Eq. (5), i.e., $\mathbf{W}^*$ and $\mathbf{p}^*$, then $\mathbf{p}^*$ is a binary vector, and $\mathrm{diag}(\mathbf{p})\mathbf{W}$ is a matrix where the elements of many rows are all zeros.
- Absorb the indicator variables $\mathbf{p}$ into $\mathbf{W}$, and use $L_{2,0}$-norm on $\mathbf{W}$ to achieve feature selection, leading to the following problem

$$\begin{aligned} \arg\min_{\mathbf{W}} \quad & \frac{1}{2}||\mathbf{X}^T\mathbf{W} - \mathbf{H}||_F^2, \\ \text{s.t.} \quad & ||\mathbf{W}||_{2,0} \leq m. \end{aligned} \tag{8}$$

- $L_{2,0}$-norm of $\mathbf{W}$ is defined as $||\mathbf{W}||_{2,0} = \mathrm{card}(||\mathbf{w}^1||_2, \ldots, ||\mathbf{w}^d||_2)$

Relaxation

- We relax $||\mathbf{W}||_{2,0} \leq m$ to its convex hull, and obtain the following relaxed problem,

$$\begin{aligned} \arg\min_{\mathbf{W}} \quad & \frac{1}{2}||\mathbf{X}^T\mathbf{W} - \mathbf{H}||_F^2, \\ \text{s.t.} \quad & ||\mathbf{W}||_{2,1} \leq m. \end{aligned} \tag{9}$$

- $L_{2,1}$-norm of $\mathbf{W}$ is defined as $||\mathbf{W}||_{2,1} = \sum_i^d ||\mathbf{w}^i||_2$
- Or equivalently the regularized problem,

$$\arg\min_{\mathbf{W}} \frac{1}{2}||\mathbf{X}^T\mathbf{W} - \mathbf{H}||_F^2 + \mu||\mathbf{W}||_{2,1}, \tag{10}$$

  where $\mu > 0$ is a regularization parameter.
- Eq. (10) can be solved by proximal gradient descent.

Outline

1. Background
2. Motivation
3. The Proposed Method
4. Experiments
5. Summary

Experimental Setting

- We use two standard face recognition databases
  - ORL face database
    - 40 persons, 10 images per person, 1024 dim
  - Extended Yale-B database
    - 38 persons, 64 images per person, 1024 dim
- For ORL (or Yale-B) data set, $p = 2, 3, 4\ (or 10, 20, 30)$ images were randomly selected as training samples for each person, and the rest images were used for testing. The training set was used to learn a subspace, and the recognition was performed in the subspace by 1-Nearest Neighbor classifier.
- Regularization parameter μ: grid search in $\{0.01, 0.05, 0.1, 0.2, 0.5\}$

Face recognition accuracy on the ORL data set

Data set	2 training		3 training		4 training	
	Acc	Dim	Acc	Dim	Acc	Dim
Baseline	66.81±3.41	–	77.02±2.55	–	81.73±2.27	–
PCA	66.81±3.41	79	77.02±2.55	119	81.73±2.27	159
FS	69.06±3.04	197	79.07±2.71	200	84.42±2.41	199
LDFS	62.69±3.43	198	75.45±2.28	192	81.96±2.56	188
LDA	71.27±3.58	28	83.36±1.84	39	89.63±2.01	39
LPP	72.41±3.17	39	84.20±1.73	39	90.42±1.41	39
FS+LDA	71.81±3.36	28	84.13±1.35	39	88.56±2.16	39
SLDA	74.14±2.92	39	84.86±1.82	39	91.44±1.53	39
LDDR	**76.88**±**3.49**	40	**86.89**±**1.91**	40	**92.77**±**1.61**	40

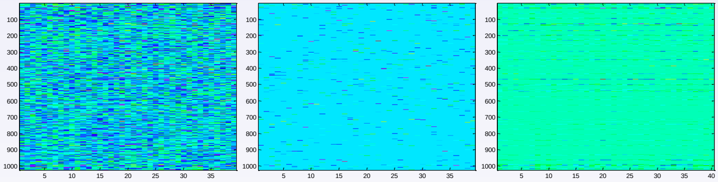
PCA: Pricinpal Component Analysis
FS: Fisher Score
LDA: Linear Discriminant Analysis
LPP: Locality Preserving Projection
FS+LDA: Fisher Score+Linear Discriminant Analysis
SLDA: Sparse Linear Discriminant Analysis
LDDR: Linear Discriminant Dimensionality Reduction

Face recognition accuracy on the Yale-B data set

Data set	10 training		20 training		30 training	
	Acc	Dim	Acc	Dim	Acc	Dim
Baseline	53.44 ± 0.82	–	69.24 ± 1.19	–	77.39 ± 0.98	–
PCA	52.41 ± 0.89	200	67.04 ± 1.18	200	74.57 ± 1.07	200
FS	64.34 ± 1.40	200	76.53 ± 1.19	200	82.15 ± 1.14	200
LDFS	66.86 ± 1.17	182	80.50 ± 1.17	195	83.16 ± 0.90	197
LDA	78.33 ± 1.31	37	85.75 ± 0.84	37	81.19 ± 2.05	37
LPP	79.70 ± 2.96	76	80.24 ± 5.49	75	86.40 ± 1.45	78
FS+LDA	77.89 ± 1.82	37	87.89 ± 0.88	37	93.91 ± 0.69	37
SLDA	81.56 ± 1.38	37	89.68 ± 0.85	37	92.88 ± 0.68	37
LDDR	$\mathbf{89.45 \pm 1.11}$	38	$\mathbf{96.44 \pm 0.85}$	38	$\mathbf{98.66 \pm 0.43}$	38

PCA: Pricinpal Component Analysis
FS: Fisher Score
LDA: Linear Discriminant Analysis
LPP: Locality Preserving Projection
FS+LDA: Fisher Score+Linear Discriminant Analysis
SLDA: Sparse Linear Discriminant Analysis
LDDR: Linear Discriminant Dimensionality Reduction

Linear Transformation Matrices



(a) LDA (b) SLDA (c) LDDR

Figure: The linear transformation matrix learned by (a) LDA, (b) SLDA ($\mu = 50$) and (c) LDDR ($\mu = 0.5$) with 3 training samples per person on the ORL database.

Each row of the linear transformation matrix of LDDR tends to be zero or nonzero simultaneously, which leads to joint feature selection and transformation.

Selected Features

Figure: Selected features (marked by blue cross) by (a) Fisher score and (b) LDDR ($\mu = 0.5$) with 3 training samples per person on the ORL database.

The features (pixels) selected by LDDR are asymmetric.
The selected pixels are mostly around the eyebrow, the boundary of eyes, nose and cheek, which are discriminative for distinguishing face images of different people.

Outline

1. Background
2. Motivation
3. The Proposed Method
4. Experiments
5. Summary

Summary

- We proposed a unified framework namely Linear Discriminant Dimensionality Reduction (LDDR) to integrate Fisher score and Linear Discriminant Analysis (LDA).
- It is able to do joint feature selection and subspace learning.
- We developed an efficient algorithms for the framework.
- Empirical experiments showed that LDDR is better than either doing Fisher score or LDA individually.
- LDDR is also better than doing Fisher score and LDA independently in two steps.

Thank You