Best Paper Candidate in Retrieval Track, Pacific-rim Conference on Multimedia, December 11-14, 2007, Hong Kong.

FADA: An Efficient Dimension Reduction Scheme for Image Classification

Yijuan Lu[1], Jingsheng Ma[2], Qi Tian[1]
[1] Department of Computer Science, University of Texas at San Antonio, TX, USA
{lyijuan, qitian}@cs.utsa.edu
[2] Institute of Petroleum Engineering, Heriot-Watt University, Edinburgh, UK,
{Jingsheng.ma}@pet.hw.ac.uk

Abstract. This paper develops a novel and efficient dimension reduction scheme--Fast Adaptive Discriminant Analysis (FADA). FADA can find a good projection with adaptation to different sample distributions and discover the classification in the subspace with naïve Bayes classifier. FADA overcomes the high computational cost problem of current Adaptive Discriminant Analysis (ADA) and also alleviates the overfitting problem implicitly caused by ADA. FADA is tested and evaluated using synthetic dataset, COREL dataset and three different face datasets. The experimental results show FADA is more effective and computationally more efficient than ADA for image classification.

Keywords: Adaptive Discriminant Analysis, Image Classification

1 Introduction

Linear discriminant analysis (LDA) [1] and Biased Discriminant Analysis (BDA) [2] are both effective techniques for feature dimension reduction. LDA assumes that positive and negative samples are from the same sources (distributions) and makes the equivalent (unbiased) effort to cluster negative and positive samples.

Compared to LDA, BDA assumes that positive samples must be similar while negative samples may be from different categories. Hence, BDA is biased towards the positive examples. It tries to find an optimal mapping that all positive examples are clustered and all negative examples are scattered away from the centroid of the positive examples. Studies have shown that BDA works very well in image retrieval especially when the size of the training sample set is small [2].

Obviously, both LDA and BDA have their own pros and cons. In addition, many applications do not fit exactly into either of the two assumptions. Hence, an Adaptive Discriminant Analysis (ADA) [3] was proposed, which merges LDA and BDA in a unified framework and offers more flexibility and a richer set of alternatives to LDA and BDA in the parametric space.

However, ADA is a parametric method. How to find good parameters is still a difficult problem for ADA. In ADA, it needs searching the whole parameter space to find the optimal one. Hence, the computational cost is very expensive and the method becomes less efficient. In addition, excessively searching also causes overfitting problem.

In this paper, we propose an efficient dimension reduction scheme for image classification, namely FADA, which stands for Fast Adaptive Discriminant Analysis. FADA overcomes the difficulties of ADA, while achieving effectiveness. The key difference between FADA and ADA lies in the adaptation method. Instead of searching parameters, FADA can directly calculate the close-to-optimal prediction very fast based on different sample distributions.

Extensive experiments on synthetic dataset, COREL and three well-known face datasets are performed to evaluate the effectiveness of FADA and compare it with ADA. Our experiments show that: (1) FADA implicitly avoids the problem encountered in ADA; (2) FADA has distinctly lower costs in time than ADA, and achieves classification accuracy that is comparable to ADA.

2 Fast Adaptive Discriminant Analysis

2.1 Adaptive Discrimiant Analysis

In 2006, Adaptive Discriminant Analysis (ADA) [3] was proposed, which merges LDA and BDA in a unified framework and offers more flexibility and a richer set of alternatives to LDA and BDA in the parametric space. ADA can find a good projection with adaptation to different sample distributions and discover the classification in the subspace with naïve Bayes classifier.

To provide a better model fitting the complex distributions for positive and negative samples, ADA finds an optimal projection.

$$W_{\text{opt}} = \arg\max_{W} \frac{|W^{\mathrm{T}}[(1-\lambda)\cdot \mathrm{S}_{\mathrm{N}\to\mathrm{P}} + \lambda \cdot \mathrm{S}_{\mathrm{P}\to\mathrm{N}}]W|}{|W^{\mathrm{T}}[(1-\eta)\cdot \mathrm{S}_{\mathrm{P}} + \eta \cdot \mathrm{S}_{\mathrm{N}}]W|} \tag{1}$$

in which

$$S_{N->P} = \sum_{i \in Negative} (\boldsymbol{x}_i - \boldsymbol{m}_P)(\boldsymbol{x}_i - \boldsymbol{m}_P)^T \tag{2}$$

$$S_{P->N} = \sum_{i \in Positive} (\boldsymbol{x}_i - \boldsymbol{m}_N)(\boldsymbol{x}_i - \boldsymbol{m}_N)^T \tag{3}$$

$$S_P = \sum_{i \in Positive} (\boldsymbol{x}_i - \boldsymbol{m}_P)(\boldsymbol{x}_i - \boldsymbol{m}_P)^T \tag{4}$$

$$S_N = \sum_{i \in Negative} (\boldsymbol{x}_i - \boldsymbol{m}_N)(\boldsymbol{x}_i - \boldsymbol{m}_N)^T \tag{5}$$

The $\boldsymbol{m}_P$ and $\boldsymbol{m}_N$ are the means of positive and negative samples, respectively. S_P (or S_N) is the within-class scatter matrix for the positive (or negative) examples. $S_{N\to P}$ (or $S_{P\to N}$) is the between-class scatter matrix from the negative (or positive) examples to the centroid of the positive (or negative) examples. The two parameters λ and η control the bias between positive and negative samples and range from $(0,0)$ to $(1,1)$. When λ and η are set to be 0 and 0, ADA recovers BDA and when λ and η are set to 0.5 and 0.5, ADA corresponds to a LDA-like projection. Alternatives to LDA and BDA can be obtained by setting parameters λ and η.

ADA has been demonstrated that it outperforms many state-of-the-art linear and nonlinear dimension reduction methods including PCA, LDA, DEM, kernel DEM (KDEM), BDA, kernel BDA (KBDA) etc. in many various applications [3].

2.2 Fast Adaptive Discriminant Analysis

Since ADA is a parametric method, parameter optimization and selection are important but difficult. i) It needs searching the whole parametric space to find the optimal projection. Its computational cost is very expensive. ii) It is hard to decide a trade-off between computational cost and accuracy. When the step-searching size is large, it will miss the global optimal value and when the size is small, it always causes overfitting problem.

In order to solve these problems, in this paper, we propose a Fast Adaptive Discriminant Analysis (FADA). Instead of searching the parametric space, FADA provides a novel and stable solution to find close-to-optimal ADA projection very fast. It saves a lot of computational cost and alleviates the overfitting problem.

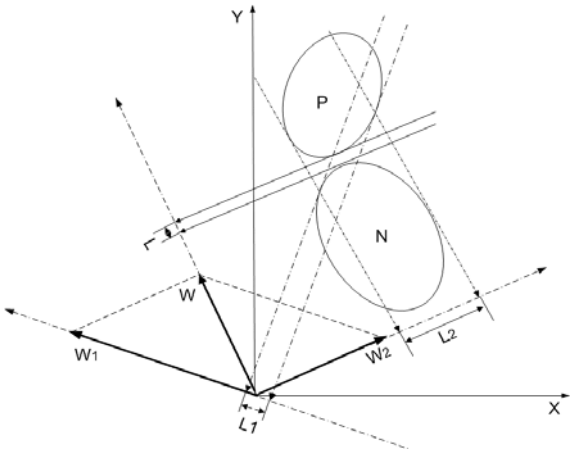


Fig.1 Illustration of FADA algorithm.

The basic idea of FADA is to find projections to cluster positive samples and negative samples respectively. Then adjust these projections to separate two classes as far as possible. Figure 1 gives an illustration of the basic idea of the FADA in two dimensional space.

The scenario can be described in the following steps (Fig. 1):

1. Firstly, find a projection W_1 that all positive samples (**P**) are clustered in the low dimensional space.
   The problem of finding the optimal W_1 can be mathematically represented as the following minimization problem:

$$W_1 = \arg\min_{W} \left| W^T S_W^P W \right| \tag{6}$$

$$S_W^P = \sum_{i=1}^{N_P} \left(\mathbf{x}_i^{(P)} - \mathbf{m}_P\right)\left(\mathbf{x}_i^{(P)} - \mathbf{m}_P\right) \quad (7)$$

Here, the within-class scatter matrix S_W^P measures the within-class variance of positive samples. $\{\boldsymbol{x}_i^{(P)}, i = 1,\ldots,N_P\}$ denote the feature vectors of positive training samples. N_P is the number of the samples of the positive class, $\boldsymbol{m}_P$ is mean vector of the positive class. Obviously, W_1 is the eigenvector(s) corresponding to the smallest eigenvalue(s) of within-class scatter matrix of positive samples.

2. Project all positive and negative data to W_1, calculate the number of samples R_1 within the overlapping range L_1 of these two classes after projection. The smaller R_1, the more separated of these two classes. If $R_1 = 0$, the positive samples and negative samples can be completely separated by the projection W_1.
3. Similarly, find a projection W_2 to cluster negative samples (**N**).

$$W_2 = \arg\min_W \left| W^T S_W^N W \right| \quad (8)$$

$$S_W^N = \sum_{i=1}^{N_R} \left(\mathbf{x}_i^{(N)} - \mathbf{m}_N\right)\left(\mathbf{x}_i^{(N)} - \mathbf{m}_N\right) \quad (9)$$

W_2 is the eigenvector(s) with the smallest eigenvalue(s) of S_W^N, within-class scatter matrix of negative samples.

4. Project all data to W_2 and calculate the number of samples R_2 belong to the overlapping range L_2 of the two classes.
5. Calculate the ratio $\lambda = \frac{R_2}{R_1 + R_2}$, $1 - \lambda = \frac{R_1}{R_1 + R_2}$
6. The final projection W is a linear combination of W_1 and W_2:

$$W = \lambda W_1 + (1 - \lambda) W_2 \quad (10)$$

Obviously, final W depends on the value of R_1 and R_2 (separability of two classes after projected by W_1 and W_2). If W_1 can better separate two classes than W_2 ($R_2 > R_1$), W will approach W_1 (W_1 has more weight). Shown in Fig. 1, after projection by the calculated W, there is no overlapping between positive samples and negative samples. Hence, in the low dimensional space, these two classes can be separated well.

3 Experiments and Analysis

3.1 FADA on Synthetic Datasets

In order to validate the effectiveness of FADA, we first use synthetic data to simulate different sample distributions as shown in Fig 2. Original data are simulated in 2-D

feature space, and positive examples are marked with "+" s and negative examples are marked with "o" s in the figure. In each case, we apply BDA, LDA and FADA to find the best projection direction by their criterion functions. The resulting projection lines are drawn in dotted, dash-dotted and solid lines, respectively. In addition, the distributions of the examples along these projections are also drawn like bell-shaped curves along projection line, assuming Gaussian distribution for each class. The thicker curves represent the distribution of projected positive examples and the thinner curves denote the distribution of projected negative examples.

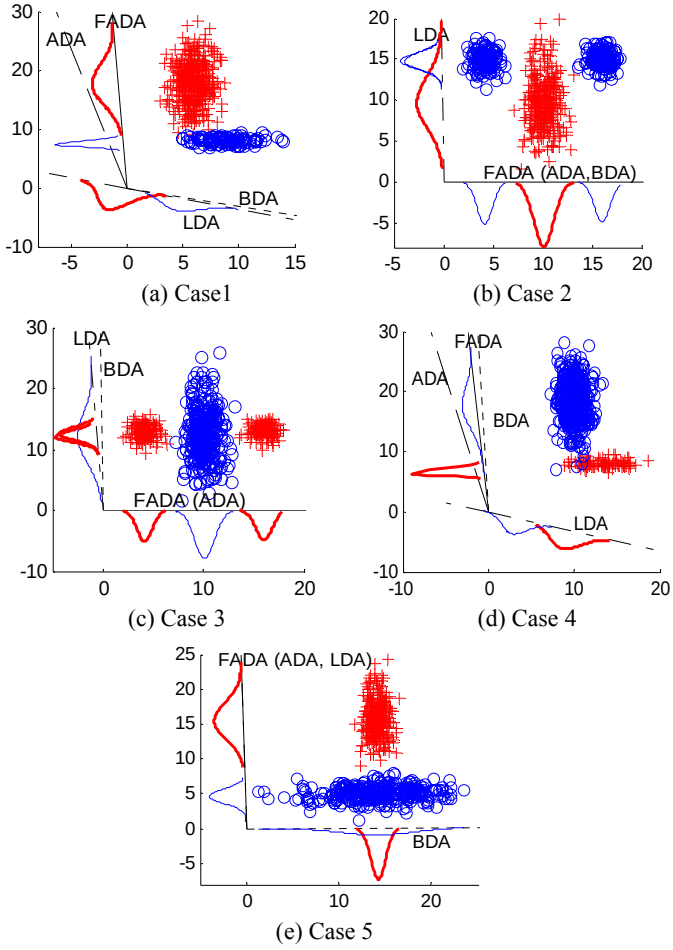


Fig.2 Comparison of optimal projections founded by LDA, BDA, ADA and FADA on synthetic data.

Shown in Fig.2, we can see these five cases actually represent several typical data distribution scenarios. Case 1 and Case 4 best represent the imbalanced data set, where the size of positive (negative) sample set is much larger than that of negative

(positive) samples (*Fig 2. (a) and (d)*). Case 2 and Case 3 best fit the distribution that the positive (negative) samples all look alike while negative (positive) ones may be irrelevant to each other and from different distributions (*Fig 2. (b) and (c)*). Case 5 is the scenario where the major descriptive directions of positive samples and negative samples are upright (*Fig 2. (e)*).

From the simulation results, we can see LDA treats positive and negative samples equally, *i.e.*, it tries to cluster the positive samples and decrease the scatter of the negative samples, even some from different sub-classes. This makes it a bad choice in Case 2 and Case 3. Similarly, since BDA assumes all positive samples are projected together, it fails in Case 3 and Case 5. In Case 1 and Case 4, BDA and LDA are found not applicable for imbalanced data sets. The reason for this is that LDA or BDA tends to severely bias to the dominating samples.

In all five cases, FADA yields as good projection as ADA with positive samples and negative samples well separated. The difference is that FADA directly computes a close-to-optimal projection easily, while ADA finds the good projection by complex and expensive parameter searching. FADA outperforms BDA and LDA as well. Note in Case 3, both BDA and LDA totally fail while FADA still produces a good projection. It clearly demonstrates that no matter if it is an imbalanced data set or samples are from different sub-classes, FADA can adaptively fit different distributions of samples fast and find a balance between clustering and separating, which are embedded in the criterion function.

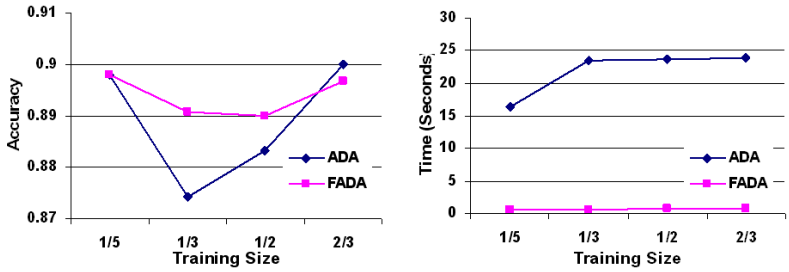
3.2 FADA for Image Classification

In section 3.2 and 3.3, we experimentally evaluate the performance of FADA on real image datasets: COREL image data set and three popular face image data sets, which cover a wide range of data in computer vision applications. The use of ADA, LDA, BDA and other state of the art methods have been investigated on the same data set [3]. The congruent results are that ADA outperformed the other algorithms with Bayesian as the base classifier. Therefore in our experiments, we focused on comparing ADA with FADA in terms of classification accuracy and efficiency (computational time). In COREL data set, ADA searches 36 parameter combinations (λ, η) sampled from 0 to 1 with step size of 0.2 to find the best one. Bayesian classifier is used on the projected data for all projection-based methods. In all experiments, average classification accuracy of 30 runs is reported. We performed our experiments using Matlab on a Pentium IV 2.26GHz machine with 1GB RAM.

In our experiments, COREL image database contains 1386 color images, which are categorized into 14 classes. Each class contains 99 images. Each image is represented by 37 feature components including color moments [4], wavelet-based texture [5] and water-filling edge-based structure features [6]. For simplicity, we randomly pick up two classes of images for classification.

Figure 3 shows the performance of ADA and FADA as the size of training samples changes from 1/5 to 2/3 of the total samples. For example, 1/5 means one-fifth of the images are used for training while the rest are used for testing. In Fig.3 (a), we find that the accuracy of our proposed FADA and ADA both change with different training sets. No matter the training size is small or large, FADA outperforms ADA in

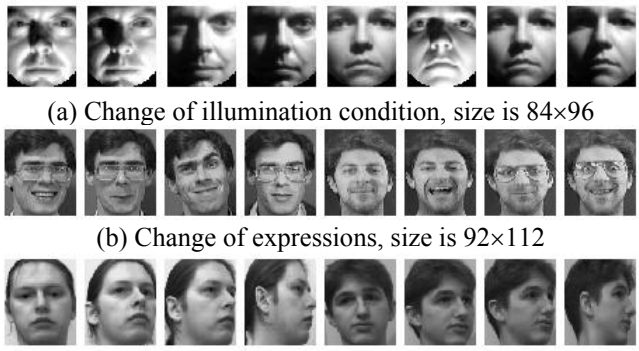
most cases or at least is comparable with ADA. Another key observation from Fig. 3 (b) is that FADA is much faster than ADA. As the size of the training set increases, the speedup of FADA over ADA significantly increases because ADA spends a lot of time in training and searching. It demonstrates that FADA is a more effective and efficient dimension reduction algorithm than ADA, as it is competitive to ADA in classification while it has much lower time costs.



(a) **(b)**

Fig.3 Comparison of accuracy and efficiency for ADA and FADA with different sizes of training set

3.3 FADA for Face Classification



(a) Change of illumination condition, size is 84×96

(b) Change of expressions, size is 92×112

(c) Change of head pose, size is 92×112

Fig.4. Example Face images from three facial databases

To evaluate FADA for face classification, we tested FADA on three well-known face image databases with change in illumination, expression and head pose, respectively. The Harvard Face image database contains images from 10 individuals, each providing 66 images, which are classified into 10 sets based on increasingly changed illumination condition [7]. The AT&T Face Image database [8] consists of grayscale images of 40 persons. Each person has 10 images with different expressions, open or closed eyes, smiling or non-smiling and wearing glasses or no glasses. The UMIST

Face Database [9] consists of 564 images of 20 people, which cover a range of poses from profile to frontal views. Figure 4 gives some example images from the databases. Sixty image features are extracted to represent these images including histogram (32), wavelet-based texture (10) and water-filling edge-based structure features (18).

For each database, we randomly chose one person's face images as positive and the rest face images of others are considered as negative. For comparison purpose, ADA, and FADA are tested on the same databases. In all of these data sets, ADA searches 121 various parameter combinations with searching step size of 0.1.

Table 1. Comparison of classification accuracy and efficiency on three different face databases

Datasets	Method	ADA	FADA
Harvard Subset1	*Accuracy (%)*	89.56	**89.67**
	Time (Second)	78.67	**0.72**
Harvard Subset2	*Accuracy (%)*	88.62	**88.70**
	Time (Second)	114.34	**0.98**
Harvard Subset3	*Accuracy (%)*	88.98	**89.58**
	Time (Second)	155.93	**1.31**
ATT Dataset	*Accuracy (%)*	**97.88**	97.28
	Time (Second)	328.77	**2.89**
UMIST Dataset	*Accuracy (%)*	95.55	**95.76**
	Time (Second)	471.56	**4.31**

Table 1 shows the comparison of ADA and FADA on accuracy and efficiency, with the largest accuracy and the smallest computational time in bold. It can be seen that FADA performs better in 4 out of 5 datasets on classification accuracy and at least two orders of magnitude faster than ADA in all 5 datasets. It is to be noted that the computation requirements of ADA increase cubically with the increase size of datasets (from Harvard to UMIST dataset) and the speed difference between ADA and FADA becomes more significant with the increase of face database scale. It is proved that FADA not only reduces the computational cost, but also increases the accuracy. It is an efficient dimension reduction scheme for image classification on small or large image datasets.

4 Conclusion and Future Work

In this paper, we propose a Fast Adaptive Discriminant Analysis (FADA) to alleviate the expensive computation cost of ADA. The novelty lies in that instead of searching a parametric space, it calculates the close-to-optimal projection automatically

according to various sample distributions. FADA has asymptotically lower time complexity than ADA, which is desirable for large image datasets, while it implicitly alleviates the overfitting problem encountered in classical ADA. All experimental results show that FADA achieves competitive classification accuracy with ADA, while being much more efficient. Extensions of FADA to high dimensional application are our future work.

Acknowledgments. This work was supported in part by Army Research Office (ARO) grant under W911NF-05-1-0404, by Department of Homeland Security (DHS) and by the San Antonio Life Science Institute (SALSI).

References

1. Ruda, R., Hart, P., Stork, D.: Pattern classification. 2nd edition, John Wiley & Sons, Inc. (2001)
2. Zhou, X., Huang, T. S.: Small sample learning during multimedia retrieval using biasMap. IEEE CVPR (2001)
3. Yu, J., Tian, Q.: Adaptive discriminant projection for content-based image retrieval. Proc. of Intl. Conf. on Pattern Recognition, Hong Kong, August (2006)
4. Stricker, M., Orengo, M.: Similarity of color images. Proceedings of SPIE Storage and Retrieval for Image and Video Databases, San Diego, CA (1995)
5. Smith, J. R., Chang, S. F.: Transform features for texture classification and discrimination in large image database. Proceedings of IEEE International Conference on Image Processing, Austin, TX (1994)
6. Zhou, X., Rui, Y., Huang, T. S.: Water-filling algorithm: a novel way for image feature extraction based on edge maps. Proceedings of IEEE International Conference on Image Processing, Kobe, Japan (1999)
7. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Trans. PAMI, Vol. 19, No. 7, July (1997)
8. Rowley, H.A., Baluja, S. Kanade, T.: Neural network-based face detection. IEEE Trans. PAMI, Vol. 20 (1998)
9. F. Samaria, F., Harter, A.: Parameterisation of a stochastic model for human face identification. IEEE Workshop on Applications of Computer Vision, Sarasota FL, December (1994)