A New Linear Discriminant Analysis Method to Address the Over-Reducing Problem

Huan Wan[2(✉)], Gongde Guo[2], Hui Wang[1], and Xin Wei[2]

[1] School of Computing and Mathematics, University of Ulster at Jordanstown, Newtownabbey, Ireland, UK
h.wang@ulster.ac.uk

[2] Key Lab of Network Security and Cryptology School of Mathematics and Computer Science, Fujian Normal University, Fuzhou, P.R. China
huanwan.mail@qq.com

Abstract. *Linear discriminant analysis* (LDA) is an effective and efficient linear dimensionality reduction and feature extraction method. It has been used in a broad range of pattern recognition tasks including face recognition, document recognition and image retrieval. When applied to fewer-class classification tasks (such as binary classification), however, LDA suffers from the *over-reducing* problem – insufficient number of features are extracted for describing the class boundaries. This is due to the fact that LDA results in a fixed number of reduced features, which is one less the number of classes. As a result, the classification performance will suffer, especially when the classification data space has high dimensionality. To cope with the problem we propose a new LDA variant, *orLDA* (i.e., *LDA for over-reducing problem*), which promotes the use of individual data instances instead of summary data alone in generating the transformation matrix. As a result orLDA will obtain a number of features that is independent of the number of classes. Extensive experiments show that orLDA has better performance than the original LDA and two LDA variants – uncorrelated LDA and orthogonal LDA.

Keywords: Dimensionality reduction · Binary classification · Linear discriminant analysis · LDA for over-reducing problem

1 Introduction

Linear discriminant analysis (LDA) is an effective and efficient method for dimensionality reduction (feature extraction). It has been successfully used in many pattern recognition problems such as face recognition [1,2], document recognition [3] and image retrieval [4,5]. It uses within-class scatter matrix S_w to evaluate the compactness within same class, and between-class scatter matrix S_b to evaluate the separation between different classes. The objective of LDA is to find an optimal transformation matrix W which minimizes the within-class scatter matrix S_w and simultaneously maximizes the between-class scatter matrix S_b.

M. Kryszkiewicz et al. (Eds.): PReMI 2015, LNCS 9124, pp. 65–72, 2015.
DOI: 10.1007/978-3-319-19941-2_7

In past two decades various improvements over the origianl LDA have been proposed in order to enhance its performance in different ways, resulting in different LDA variants. These LDA variants can be put into two categories. In the first category, the LDA variants attempt to tackle the singularity problem of within-class scatter matrix (S_w). In LDA, we take the leading eigenvectors of $S_w^{-1}S_b$ as the columns of optimal transformation matrix W. In order to guarantee S_w nonsingular, it requires at least $N + C$ samples [6], where N is data dimension and C is the number of classes. However, in realistic world it does not always happen and it is almost impossible in high-dimensionality space. Therefore, sigularity makes within-class scatter matrix irreversible and we can not use $S_w^{-1}S_b$ to obtain transformation matrix W. In order to address the singularity problem, Li-Fen Chen et al. [1] proposed NLDA, which is short for null space linear discriminant analysis. It is based on a new Fisher's criterion function and calculates the transformation matrix in the null space of the within-class scatter matrix, which avoids the singularity problem implicitly. In [7] regularized linear discriminant analysis (RLDA) is proposed. It gets optimal constant α by heuristic approach and adds α to the diagonal elements of the within-class scatter matrix to overcome the singularity problem. Some new approaches are proposed to solve the singularity problem recently. For example, Alok Sharma et al. [8] proposed a new method to compute the transformation matrix W, which gave a new perspective to NLDA and presented a fast implementation of NLDA using random matrix multiplication with scatter matrices; Alok Sharma et al. [9] proposed an improvement of RLDA, which presented a recursive method to compute the optimal parameter; and Xin Shu et al. [10] proposed LDA with spectral regularization to tackle the singularity problem. Other LDA variants for solving the singularity problem can be found in [11].

In the second category, the LDA variants apply the original LDA in local data space instead of whole data space. For example, Zizhu Fan et al. [12] presented two local linear discriminant analysis (LLDA) approaches: vector-based LLDA (VLLDA) and matrix-based LLDA (MLLDA), which select a proper number of nearest neighbors of a test sample from a training set to capture the local data structure and use the selected nearest neighbors of the test sample to produce the local linear discriminant vectors or matrix. Chao Yao et al. [13] proposed a subset method for improving linear discriminant analysis, which divided the whole set into several subsets and used the original LDA in each subset. There are other LDA variants such as nonparametric discriminant analysis [14], sparse discriminant analysis [15], semi-supervised linear discriminant analysis [16], incremental LDA [17], tensor-based LDA [18], and local tensor discriminant analysis [19].

The original LDA and most of its variants have elegant mathematical properties, one of which being that the dimensionality of the data space can be reduced to at most one less the number of classes. One consequence is that if there are few classes in a data set, e.g., two classes in a binary classification problem, there will be one or only a few features left after the dimensionality reduction, probably insufficient for deciding the class boundaries. This leads to the *over-reducing* problem, meaning that dimensionality reduction is over done.

In this paper we propose changes to the original LDA to address the over-reducing problem. Instead of using only the means of each class and the whole data to evaluate the separation between different classes, our new LDA variant uses a new method to compute the between-class scatter matrix. As a result we get more between-class information and more features (than before) after dimensionality reduction even for binary classification.

The rest of the paper is organized as follows. Section 2 reviews the original LDA and two well known LDA variants – Uncorrelated LDA and Orthogonal LDA. Section 3 presents our orLDA, Sect. 4 presents our experimental results and Sect. 5 concludes the paper.

2 Linear Discriminnant Analysis

2.1 The Original LDA

LDA has been widely used for dimensionality reduction and feature extraction. In the original LDA, the within-class scatter matrix and the between-class scatter matrix are used to measure the class compactness and separability respectively. They are defined as [20]:

$$S_w = \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{n_i}(x_{ij}-\mu_i)(x_{ij}-\mu_i)^T, \tag{1}$$

$$S_b = \frac{1}{N}\sum_{i=1}^{C} n_i(\mu_i-\mu)(\mu_i-\mu)^T, \tag{2}$$

where N denotes the number of data samples, C denotes the number of the classes, n_i denotes the number of samples in class i, μ_i denotes the mean of samples in class i, μ denotes the mean of whole samples, and x_{ij} is the jth sample in class i. The original LDA aims to find a transformation matrix $W_{opt} = [w_1, w_2, ..., w_f]$ that maximizes the Fisher's criterion

$$J(W) = \frac{W^T S_b W}{W^T S_w W} \tag{3}$$

Mathematically, the solution to this problem corresponds to an eigenvalue decomposition of $S_w^{-1}S_b$, taking its leading eigenvectors as the columns of W_{opt}.

From Eq. (2), we can see that LDA uses only the centers of classes and whole data set to compute between-class scatter matrix. This may lose much class-separating information. Because the rank of the between-class matrix is at most $C-1$, the number of extracted features by LDA is at most $C-1$. However, it is insufficient to separate the classes well with only $C-1$ features, especially for binary classification in high-dimensional spaces.

2.2 Uncorrelated LDA and Orthogonal LDA

Uncorrelated LDA (ULDA) and Orthogonal LDA (OLDA) were presented in [21]. In this paper, Jieping Ye proposed a new optimization criterion to obtain the optimal transformation matrix W_{opt}. W_{opt} is defined as: $W_{opt} = X_q M$, where X is a matrix that simultaneously diagonalizes S_b, S_w, S_t[1], X_q is the matrix consisting of the first q columns of X, and M is an arbitrary nonsingular matrix. When M is the identity matrix, we can get Uncorrelated LDA algorithm and make features in the reduced space uncorrelated; however, if we let X_q=QR be the QR decomposition of X_q and choose M as the inverse of R, we get Orthogonal LDA algorithm and make the discriminant vectores of OLDA orthogonal to each other.

3 Linear Discriminnant Analysis that Avoids Over-Reducing

In this section we present the proposed changes to the original LDA in order to address the over-reducing problem, which are related to how to compute the between-class matrix.

Suppose there are N samples $X_i \in R^n$ for $i = 1, 2, \ldots, N$ from two classes, N_k is the number of samples in class k $(k = 1, 2)$ such that $\sum_{k=1}^{2} N_k = N$, μ_k is the mean of the samples in class k, and x_{kj} is the jth sample in class k. Two scatter matrices, the within-class scatter matrix $(\widetilde{S_w})$ and between-class scatter matrix $(\widetilde{S_b})$ are defined as follows:

$$\widetilde{S_w} = \frac{1}{N} \sum_{k=1}^{2} \sum_{j=1}^{N_k} (x_{kj} - \mu_k)(x_{kj} - \mu_k)^T \tag{4}$$

$$\widetilde{S_b} = \frac{1}{N} (N_1 \sum_{j=1}^{N_1} (x_{1j} - \mu_2)(x_{1j} - \mu_2)^T + N_2 \sum_{j=1}^{N_2} (x_{2j} - \mu_1)(x_{2j} - \mu_1)^T) \tag{5}$$

When the number of classes is two, Eq. (1), the computation of within-class scatter matrix in the original LDA, is the same as Eq. (4). However, Eq. (5), the computation of between-class scatter matrix, is quite different from the original LDA. In Eq. (5), we use every sample in one class to subtract the mean of another class.

It is clear that computing the between-class scatter matrix in this way will capture more between-class information than the original LDA hence we can expect better classification performance. Besides, by Eq. (5), we can get more than 1 feature in binary classification. According to linear algebra and Eqs. (4) and (5), we can obtain $rank(S_b) = min(n, N-2)$ and $rank(S_w) = min(n, N-2)$, where n is the dimensionality of data space and N is the total number of samples.

[1] S_t denotes total scatter matrix, which is defined as: $S_t = \frac{1}{N} \sum_{j=1}^{N} (x_j - \mu)(x_j - \mu)^T$, where x_j denotes the jth sample.

Then we can find that the ranks of S_b and S_w depend only on n and N. Therefore, $rank(S_w^{-1}S_b) = min(rank(S_w^{-1}), rank(S_b))$ is not limited by 1 extracted feature. Our optimal transformation matrix W_{opt} maximizes $J(W) = \frac{W^T \widetilde{S_b} W}{W^T \widetilde{S_w} W}$ and we get the eigenvectors corresponding to the top eigenvalues of the eigenequation $\widetilde{S_w}^{-1}\widetilde{S_b}$ as columns of W_{opt}.

4 Experiments

In this section we take *K-Nearest Neighbor* (KNN, K=1) as the classifier and use ten-fold cross-validation to evaluate our method on three face datasets – ORL face database[2], Labeled Faces in the Wild (LFW) [22], and Extended Cohn-Kanade [23]; and one DNA microarray gene expression datasets from Kent Ridge Bio-medical Dataset (KRBD)[3].

The ORL face database consists of a total of 400 images of 40 distinct people. Each people has ten different images and the size of each image is 92*112 pixels, with 256 grey levels per pixel. All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position.

LFW face dataset consists of 13,233 images of 5,749 people, which are organized into 2 views – a development set of 3,200 pairs for building models and choosing features; and a ten-fold cross-validation set of 6,000 pairs for evaluation. The size of each image is 250*250 pixels. All the images are collected from the Internet with large intra-personal variations. There are three versions of the LFW: original, funneled and aligned. In our experiment, we use the aligned version [24].

For the above two face datasets, we do face verification experiment, which is a binary classification problem. The goal of face verification is to decide if two given face images match or not. We use subset of view2 of LFW. We randomly choose 200 matched face pairs and 200 mismatched face pairs from view2 and crop each image to an image of 80 * 150 pixels as in [25]. However, for ORL face dataset, through randomly matching face images, we obtain 80 matched face pairs and 391 mismatched face pairs for face verification. Therefore, we have 400 samples of LFW and 471 samples of ORL. The dimensionality of each sample in LFW and ORL are 24,000 and 20,608, respectively.

Extended Cohn-Kanade dataset (CK+) is a complete dataset for action unit and emotion-specified expression. In this paper, we focus on emotion-specified expressions. There are 593 sequences from 123 subjects which are FACS coded at the peak frame, but only 327 of the 593 sequences have emotion sequences and use the last frame of each sequence to do expression classification. There are seven kinds of emotion expression, including: neutral, anger, contempt, disgust, fear, happy, sadness and surprise. Here, we do positive and negative expression classification experiment and take happy as positive expression and the rest of emotion as negative expression. Therefore, we have 69 positive expression

[2] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.
[3] http://datam.i2r.a-star.edu.sg/datasets/krbd/.

samples and 258 negative expression samples and the dimensionality of each sample is 10,000.

Acute Leukemia dataset [26] consists of DNA microarray gene expression data of human acute leukemia data for cancer classification. There are two types of acute leukemia: 47 acute lymphoblastic leukemia (ALL) and 25 acute myeloid leukemia (AML), over 7129 probes from 6817 human genes.

We compare our orLDA with three discriminant dimension reduction methods, which are the original LDA, Uncorrelated LDA (ULDA) and Orthogonal LDA (OLDA) [21]. To guarantee that S_w does not become singular, we use two-stage PCA+LDA [27] – we reduce the data dimensionality by PCA, retaining principal components which explain 95 % of variance, before original LDA and orLDA methods are used.

Experimental results on the four datasets are shown in Table 1. It is clear that our orLDA has better classification performance than the original LDA, ULDA and OLDA on all datasets except Extended Cohn-Kanade. We credit the better performance to the facts that (1) orLDA obtains more between-class information than the other three LDA variants; (2) more than 1 extracted features can better separate two classes.

Table 1. Mean accuracy and standard error of the mean on four datasets

Datasets	ORL	LFW	Extended Cohn-Kanade	Acute Leukemia
Original LDA	0.8536 ± 0.0103	0.5675 ± 0.0190	0.9695 ± 0.0101	0.9857 ± 0.0143
orLDA	0.8832 ± 0.0102	0.625 ± 0.0194	0.9695 ± 0.0101	0.9857 ± 0.0143
ULDA	0.7684 ± 0.0179	0.58 ± 0.0244	0.9757 ± 0.0099	0.9589 ± 0.0299
OLDA	0.7684 ± 0.0179	0.58 ± 0.0244	0.9757 ± 0.0099	0.9589 ± 0.0299

5 Conclusion

In this paper, we propose a new LDA, orLDA, to address the over-reducing problem associated with LDA. orLDA uses a new method to compute between-class scatter matrix, which contains more between-class information and allows extracting more features. Experiments have shown that orLDA outperformed the original LDA, ULDA and OLDA significantly on two face datasets, outperformed ULDA and OLDA on the gene expression dataset. orLDA achieved the same performance as the original LDA on the emotion expression dataset and the gene expression dataset, and underperformed ULDA and OLDA slightly on the emotion expression dataset. It is then reasonable to conclude that the new LDA variant is an improvement over the state of the art.

References

1. Chen, L.-F., Liao, H.-Y.M., Ko, M.-T., Lin, J.-C., Yu, G.-J.: A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recognit. **33**(10), 1713–1726 (2000)
2. Zhao, X., Evans, N., Dugelay, J.: Semi-supervised face recognition with LDA self-training. In: 2011 18th IEEE International Conference on Image Processing (ICIP), pp. 3041–3044. IEEE (2011)
3. He, C.L., Lam, L., Suen, C.Y.: Rejection measurement based on linear discriminant analysis for document recognition. Int. J. Doc. Anal. Recognit. (IJDAR) **14**(3), 263–272 (2011)
4. Swets, D.L., Weng, J.J.: Using discriminant eigenfeatures for image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **18**(8), 831–836 (1996)
5. He, X., Cai, D., Han, J.: Learning a maximum margin subspace for image retrieval. IEEE Trans. Knowl. Data Eng. **20**(2), 189–201 (2008)
6. Martínez, A.M., Kak, A.C.: PCA versus LDA. IEEE Trans. Pattern Anal. Mach. Intell. **23**(2), 228–233 (2001)
7. Guo, Y., Hastie, T., Tibshirani, R.: Regularized linear discriminant analysis and its application in microarrays. Biostatistics **8**(1), 86–100 (2007)
8. Sharma, A., Paliwal, K.K.: A new perspective to null linear discriminant analysis method and its fast implementation using random matrix multiplication with scatter matrices. Pattern Recognit. **45**(6), 2205–2213 (2012)
9. Sharma, A., Paliwal, K.K., Imoto, S., Miyano, S.: A feature selection method using improved regularized linear discriminant analysis. Mach. Vis. Appl. **25**(3), 775–786 (2014)
10. Shu, X., Lu, H.: Linear discriminant analysis with spectral regularization. Appl. Intel. **40**(4), 724–731 (2014)
11. Ye, J., Ji, S.: Discriminant analysis for dimensionality reduction: An overview of recent developments. In: Boulgouris, N.V., Plataniotis, K.N., Micheli-Tzanakou, E. (eds.) Biometrics: Theory, Methods, and Applications. Wiley-IEEE Press, New York (2010)
12. Fan, Z., Xu, Y., Zhang, D.: Local linear discriminant analysis framework using sample neighbors. IEEE Trans. Neural Netw. **22**(7), 1119–1132 (2011)
13. Yao, C., Lu, Z., Li, J., Xu, Y., Han, J.: A subset method for improving linear discriminant analysis. Neurocomputing **138**, 310–315 (2014)
14. Li, Z., Lin, D., Tang, X.: Nonparametric discriminant analysis for face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **31**(4), 755–761 (2009)
15. Clemmensen, L., Hastie, T., Witten, D., Ersbøll, B.: Sparse discriminant analysis. Technometrics **53**(4), 406–413 (2011)
16. Zhao, M., Zhang, Z., Chow, T.W., Li, B.: A general soft label based linear discriminant analysis for semi-supervised dimensionality reduction. Neural Netw. **55**, 83–97 (2014)
17. Pang, S., Ozawa, S., Kasabov, N.: Incremental linear discriminant analysis for classification of data streams. IEEE Trans. Syst. Man Cybern. Part B: Cybern. **35**(5), 905–914 (2005)
18. Li, M., Yuan, B.: 2D-LDA: a statistical linear discriminant analysis for image matrix. Pattern Recognit. Lett. **26**(5), 527–532 (2005)
19. Nie, F., Xiang, S., Song, Y., Zhang, C.: Extracting the optimal dimensionality for local tensor discriminant analysis. Pattern Recognit. **42**(1), 105–114 (2009)
20. Webb, A.R.: Statistical Pattern Recognition. Wiley, New York (2003)

21. Ye, J.: Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. J. Mach. Learn. Res. **6**, 483–502 (2005)
22. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report, Technical Report 07–49, University of Massachusetts, Amherst (2007)
23. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 94–101. IEEE (2010)
24. Wolf, L., Hassner, T., Taigman, Y.: Similarity scores based on background samples. In: Zha, H., Taniguchi, R., Maybank, S. (eds.) ACCV 2009, Part II. LNCS, vol. 5995, pp. 88–97. Springer, Heidelberg (2010)
25. Kan, M., Xu, D., Shan, S., Li, W., Chen, X.: Learning prototype hyperplanes for face verification in the wild. IEEE Trans. Image Process. **22**(8), 3310–3316 (2013)
26. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286**(5439), 531–537 (1999)
27. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 711–720 (1997)