Dimensionality Reduction by Minimal Distance Maximization

Bo Xu, Kaizhu Huang, Cheng-Lin Liu
National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, 95 Zhongguancun East Road Beijing 100190, P.R. China
Email:{box, kzhuang, liucl}@nlpr.ia.ac.cn

Abstract

In this paper, we propose a novel discriminant analysis method, called Minimal Distance Maximization (MDM). In contrast to the traditional LDA, which actually maximizes the average divergence among classes, MDM attempts to find a low-dimensional subspace that maximizes the minimal (worst-case) divergence among classes. This "minimal" setting solves the problem caused by the "average" setting of LDA that tends to merge similar classes with smaller divergence when used for multi-class data. Furthermore, we elegantly formulate the worst-case problem as a convex problem, making the algorithm solvable for larger data sets. Experimental results demonstrate the advantages of our proposed method against five other competitive approaches on one synthetic and six real-life data sets.

1. Introduction

Linear discriminant analysis (LDA) [2], one of the most popularly used discriminant analysis algorithms, has been widely used in various fields including economics, psychology, neuroscience, bio-informatics, and computer science. The principle of LDA is to maximize the between-class covariance while minimizing the within-class covariance. Under the homoscedastic Gaussian assumption, LDA leads to the optimal projection axis used for two-category data. When used for multi-class data, its performance is still workable in many cases. Fig. 1 (a) is one example where LDA can find the satisfied projection axis for class separation among three classes. However, LDA may fail to find good projection for multi-class data [4]. Fig. 1 (b) illustrates a typical example. Clearly, by LDA, the data of class 1 and class 2 will be mixed with each other, leading to worse performance for consequent classification. This problem of LDA is called the class separation problem in the literature [8]. In contrast, the axis proposed by our method, the dashed axis in Fig. 1 (b), would be a reasonable projection axis that can appropriately make the data of each class well separated.

The class separation problem of LDA for multi-class data roots in the fact that LDA exploits an *average* setting, i.e., LDA tries to maximize the *average* divergences [1] among different classes.[2] To maximize the *average* divergence, LDA tends to find the subspace preserving the larger divergences and ignoring the smaller divergences, as illustrated in Fig. 1 (b). This causes the overlap of the similar classes, with smaller divergences, after data transformation.

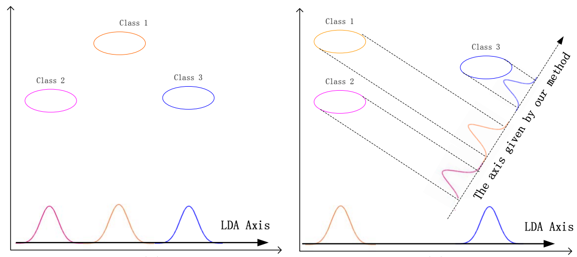
To overcome the class separation problem of LDA, in this paper, a novel *worst-case* framework called Minimal Distance Maximization (MDM) is proposed. More specifically, instead of maximizing the *average* divergence among different classes, MDM attempts to maximize the *minimal* (worst-case) divergence. In this worst-case setting, MDM tries to push away each pair of classes with small divergence instead of making the average divergence among all the classes as large as possible. This will consequently avoid the aforementioned problem. Obviously, the proposed MDM method is still optimal for two-class problems under the homoscedastic Gaussian assumption, since it is degraded to the standard LDA when the class number is equal to two. Hence, the proposed worst-case method can be seen as a more generalized version of LDA for multi-class problems. Furthermore, we elegantly transform the MDM optimization problem to a convex programming problem, making the model solvable for large data sets.

In the literature, there are many dimensionality reduction approaches related to our work [4, 5, 7, 3, 6, 10, 8]. Sugiyama [6] developed the Local Fisher Discriminant Analysis (LFDA) method that combines the merits of Locality Preserving Projection (LPP) [3] into LDA. However, its performance is limited in handling the class separation problem. Loog [4] presented an extended criterion named approximate Pairwise Accu-

[1] The divergence of any two classes is defined as the distance between the mean vectors of the two classes in the whitening space.

[2] We prove this in Lemma 1.

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.144



(a) (b)

Figure 1. Illustration of LDA for multi-class data

racy Criterion (aPAC) that adds weights in the estimation of the between-class covariance. Although this method is well motivated and can also partially solve the class separation problem, it remains a problem how to select an optimal weighting function. Another related approach is called Fractional-Step linear discriminant analysis (F-LDA) [5]. It improved the robustness of choosing the weighting function by reducing dimensionality iteratively. However, the large-scale iteration and the choice of a scaling parameter limit its application. Marginal Fisher Analysis (MFA) is also highly related to our method. However, it involves a time-consuming graph construction step. There are still many other approaches related to our work including the methods proposed in [7, 3, 10, 8]. For space limitation, we will not elaborate them here.

The paper is organized as follows. In the next section, we present our novel worst-case framework for dimensionality reduction in detail. In Section 3, we evaluate our algorithm and report experimental results. Finally, we set out the conclusion with some remarks.

2. Minimal Distance Maximization for Dimensionality Reduction

In this section, we first present the notation and then describe our novel dimension reduction framework MDM.

2.1. Notation and Two-step View of LDA

We first present the notation used in the paper. Let $x_i \in R^D$ $(i = 1, 2, ...n)$ be D-dimensional samples and $c_i \in \{1, 2, ...\ell\}$ be their associated class labels, where n is the number of samples and ℓ is the number of classes. Let n_i be the number of samples in the class i, $\sum_{i=1}^{\ell} n_i = n$. Let $X = [x_1, x_2, ...x_n]$ represent all samples as a matrix. The purpose of linear dimensionality reduction is to find a projection matrix W which maps a D-dimensional data space to a d-dimensional subspace ($d < D$), $W^T : R^D \rightarrow R^d$, where $W = [w_1, w_2, ...w_d]$. Let $Y = [y_1, y_2, ...y_n]\,, y_i \in R^m$ represent all samples in the embedding space projected by matrix W^T, where y_i is given by $y_i = W^T x_i, i \in [1, n]$. Let m_i, m'_i and M_i be the mean of the class i in the original dimensional space, the whitening space, and the low dimensional space respectively. In addition, $A \succeq 0$ means the matrix A is a semi-definite matrix.

Before we describe our novel MDM framework, we present a two-step view of LDA [2]. The transformation matrix W of LDA is usually given by the eigenvalue decomposition of $S_W^{-1} S_B$, where S_W is the within-class covariance and S_B is the so-called between-class covariance. The solution of LDA can be equivalently computed in two steps by whitening S_W first and then conducting Principle Component Analysis (PCA) in the whitening space.

Denote the eigenvectors of S_W by the matrix P and the eigenvalues by the matrix Λ_1, then we have $S_W = P \Lambda_1 P^T$. The first step of LDA is to transform S_W to an identity matrix using a whitening transformation matrix $W_1 = P \Lambda_1^{-1/2}$, i.e., $W_1^T S_W W_1 = I$. Accordingly, S_B is transformed to $W_1^T S_B W_1 = S'_B$. The second step of LDA is to utilize the PCA transformation matrix W' on class centers. This is equivalent to the transformation that maximizes the average divergence among all the classes in the whitening space. It is proved in Lemma 1. Thus the final transformation of LDA is the combination of the two step, the whitening transformation W_1 and the PCA transformation W':
$W = W_1 W' = P \Lambda_1^{-1/2} W'$.

Lemma 1 *The LDA solution is a linear transformation maximizing the average divergence among all the classes in the whitening space.*

Proof: Without consideration of the class prior probability, LDA criterion can be rewritten as [4]:

$$J_F(W) = \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} tr\left(\left(W^T S_W W \right)^{-1} \left(W^T S_{ij} W \right) \right),$$

where $S_{ij} = (m_i - m_j)(m_i - m_j)^T$. Since the within-class scatter matrix S_W is equal to I_n in the whitening space, $J_F(W)$ can be stated as follows:

$$J_F(W') = \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} tr\left(W'^T (m_{i'} - m_{j'}) (m_{i'} - m_{j'})^T W' \right) = \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \|M_i - M_j\|^2 .$$

Hence the criterion of LDA is equivalent to maximizing the average divergence among all the classes in the whitening space. This completes the proof. □

Maximizing the average divergence will possibly lead to serious overlap of those similar classes (as illustrated in Fig. 1). In the next subsection, we show our proposed worst-case framework MDM can solve this problem.

2.2. The MDM Algorithm

To attack the class separation problem, we present a novel method called MDM that maximizes the minimal pairwise divergence among the class centers. In this worst-case setting, the closer classes will be pushed further away and this hence alleviates the overlap problem systematically.

Our approach, MDM, also follows the two-step framework as LDA. After applying the whitening transformation on the data, it finds a projection matrix W' to satisfy the following criterion:

$$\max_{W'} \left(\min_{i,j} \|\mathrm{M}_i - \mathrm{M}_j\|^2 \right). \tag{1}$$

The major difference from LDA is that, MDM is maximizing the minimal divergence, while LDA is maximizing the average divergence(as proved by Lemma 1). Maximizing the average divergence may lead some classes overlap in the transformed space, while maximizing the minimal divergence can avoid such cases effectively.

We now show how to solve the optimization problem of MDM.

Eq. (1) is equivalent to:

$$\max_{W'} y \qquad \text{s.t.} \quad D_{ij} \geq y, \forall i, j \tag{2}$$

where $D_{ij} = \|\mathrm{M}_i - \mathrm{M}_j\|^2$. As D_{ij} can be rewritten as

$$D_{ij} = tr\left(W'^T \left(m'_i - m'_j \right) \left(m'_i - m'_j \right)^T W' \right).$$

Using matrix property, $tr(AB) = tr(BA)$, we obtain

$$D_{ij} = tr\left(W' W'^T \left(m'_i - m'_j \right) \left(m'_i - m'_j \right)^T \right).$$

Hence we can rewrite Eq. (2) as: $tr\left(AS'_{ij}\right) \geq y, \forall i, j$, where $A = W'W'^T$ and S'_{ij} is the scatter matrix between class i and class j in the whitening space.

After adding a reasonable constraint to matrix A for avoiding trivial solution, i.e., $\|A\| <= 1$, our criterion can be rewritten as:

$$\max_{A} \quad y \quad \text{s.t.} \quad \begin{cases} tr\left(AS'_{ij}\right) \geq y, \forall i, j \\ \|A\| <= 1 \\ A \succeq 0 \end{cases}. \tag{3}$$

The problem of (3) forms a typical Semi-Definite Programming (SDP) problem [9] that can be easily solved by many software packages, e.g. the CVX package.[3] In real implementation, the solution of (3) is usually semi-definite without the constraint of $A \succeq 0$. Hence the resulting optimization problem can be simplified to a quadratic programming problem after removing the constraint of $A \succeq 0$, which is much easier to be solved.

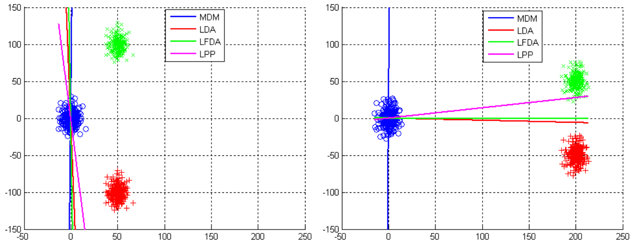


Figure 2. Projection directions of different methods

Once we obtain matrix A, the next problem is how to compute W' from A. Therefore, we define following criterion to solve W':

$$J\left(W'\right) = \min_{W'} \left\| A - W'W'^T \right\|^2. \tag{4}$$

It is easy to prove that the optimal W' is the first d largest eigenvectors of A.

Therefore, the final transformation of MDM is the combination of the whitening transformation W_1 and W': $W = W_1 W'$.

3. Experimental Results

In this section, we compare the performance of MDM with five competitive methods, LDA, LPP [3], LFDA [6], MFA [10] and aPAC [4] on one synthetic and six real data sets.

3.1. Results on Synthetic Data

To test the performance obtained by MDM, 250 samples for each of the three classes (750 samples in total) are generated in our experiment. Moreover, the samples in each class are obtained from a two-dimensional Gaussian with standard normal distribution. Each approach finds a projection axis to separate different classes as much as possible. The projection axes are shown in Fig. 2. In the left sub-figure of Fig. 2, all the approaches successfully separate different classes. However, only MDM can separate classes correctly in the right sub-figure of Fig. 2, while the other methods tend to merge closer classes. This clearly shows the superiority of MDM over the other approaches for multi-class problems.

3.2. Results on Real Data

We also evaluate our method on Iris, Opt-digits, Statlog-Satimage and Balance multi-category data sets

[3] http://stanford.edu/ boyd/cvx/download.html

Table 1. Classification rate on UCI data (mean ± std -dev%)

Data Sets	Iris	Statlog	Opt-digits	Balance
MDM	**98.0±4.5**	82.1±3.8	**95.6±1.2**	**91.7±3.8**
LDA	**98.0±4.5**	79.4±5.1	93.6±1.3	91.5±4.0
aPAC	**98.0±4.5**	**82.5±3.8**	**95.6±1.2**	91.5±4.0
LFDA	96.0±6.4	81.9±3.0	93.5±1.0	91.5±3.8
LPP	92.0±6.9	77.5±3.2	94.4±1.1	77.0±7.9
MFA	94.0±7.3	68.8±8.9	75.0±3.5	86.4±6.2

from the UCI machine learning repository.[4] For simplicity, we define the reduced dimensionality as the class number minus 1 in the experiments on UCI data. The reported accuracy is acquired by the linear SVM classifier using 10-fold Cross Validation (CV) after the data are transformed based on projections obtained by the various methods.[5] Experimental results are listed in Table 1. As can be seen from Table 1, MDM performs competitive to aPAC, while it outperforms the other methods significantly. As discussed before, aPAC is also well justified for solving the class separation problem. However, it needs to define weighting functions beforehand, which is somehow ad-hoc. In contrast, MDM presents a more systematic approach to handle the class separation problem.

To evaluate the separability on different dimensionality of MDM, we further conduct experiments on the Yale[6] and the ORL[7] face data sets. The Yale face data set contains 165 gray scale images of 15 individuals, each individual has 11 images. The ORL face data set consists of ten different images of each of 40 distinct subjects. All the face images are manually aligned and cropped. The size of each cropped image is 16 × 16 pixels to save the experimental time, with 256 gray levels per pixel. We use 8 images per individual for training and the rest for testing on both the Yale and ORL data sets. We set the dimensionality of the low dimensional subspace from 1 to the class number minus 1. Again, the linear SVM is employed as the classifier. The recognition accuracy (obtained using 10-fold CV) is presented in Fig. 3 . Clearly observed, although MDM performs closely to other approaches on Yale, it indeed outperforms the others including aPAC on ORL. This once again validates the effectiveness of our proposed method.

4. Conclusion

We proposed a novel dimensionality reduction algorithm, MDM, based on the maximization of the minimal divergence among the different classes. This solved the class separation problem of LDA. Experiments in-

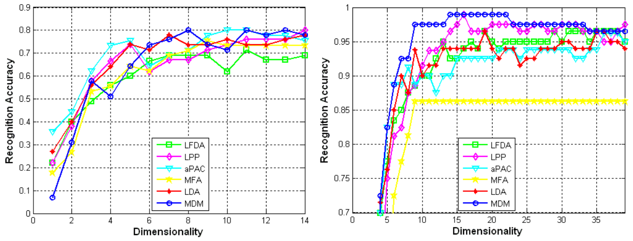


Figure 3. Comparisons on Yale (Left) and ORL (Right)

dicated the superiority of MDM over previous methods such as LDA, LFDA, and LPP.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under grant no.60825301.

References

[1] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[2] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[3] X. He and P. Niyogi. Locality preserving projections. *Advances in Neural Information Processing Systems*, 16:153–160, 2003.

[4] M. Loog, R. Duin, and R. Haeb-Umbach. Multiclass linear dimension reduction by weighted pairwise Fisher criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(7):762–766, 2001.

[5] R. Lotlikar and R. Kothari. Fractional-step dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):623–627, 2000.

[6] M. Sugiyama. Local Fisher discriminant analysis for supervised dimensionality reduction. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 905–912. ACM, 2006.

[7] E. Tang, P. Suganthan, X. Yao, and A. Qin. Linear dimensionality reduction using relevance weighted LDA. *Pattern Recognition*, 38(4):485–493, 2005.

[8] D. Tao, X. Li, X. Wu, and S. Maybank. Geometric mean for subspace selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):260–274, 2009.

[9] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, 1996.

[10] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, 2007.

[4] http://archive.ics.uci.edu/ml/datasets.html

[5] In our experiment, Libsvm [1] is adopted.

[6] http://cvc.yale.edu/projects/yalefaces/yalefaces.html

[7] http://www.orl.co.uk/facedatabase.html