On Solving the Small Sample Size Problem for Marginal Fisher Analysis

Fadi Dornaika[1,2] and Alireza Bosagzadeh[1]

[1] University of the Basque Country UPV/EHU, San Sebastian, Spain
[2] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

Abstract. Marginal Fisher Analysis (MFA) was introduced to remedy some of the shortcomings of the Fisher Discriminant Analysis (FDA). It performs local discrimination between classes. Whenever the training data set is small, MFA cannot directly be used with the original high-dimensional samples. This is referred to as the small sample size (SSS) phenomenon that happens whenever the feature dimension is higher than the number of examples. The classic remedy was using the projection of the raw data (e.g., using (PCA)). This paper introduces two regularization schemes that overcome the singularity and near singularity of the locality preserving scatters. The first scheme uses ridge regression regularization. The second scheme uses matrix exponential and introduces an implicit distance diffusion mapping. The experiments are conducted on four face data sets. These experiments demonstrate that the introduced schemes can enhance the performance of the MFA framework much better than the widely used PCA based regularization.

1 Introduction

The linear Manifold Learning paradigms are more and more used in data mining and machine learning [1]. These methods provide an explicit embedding from high dimensional space into latent spaces having lower dimension. These approaches can enhance the classification performance. The classic linear approaches (e.g., PCA, FDA, Maximum Margin Criterion (MMC)[2]) are suitable for many tasks, such as classification and recognition. PCA embeds the data samples using projection axes having the maximal variances. Unlike PCA which is a unsupervised technique, FDA [3] is supervised and seeks axes that enhance data discrimination. Several linear approaches for dimensionality reduction can be obtained from a data graph where the samples are the nodes and the similarity between samples are encoded by the edges. [4] proposes a supervised technique called average Neighborhood Margin Maximization (ANMM). In this method, the authors seek a linear embedding that maximizes the sum of margin distances (computed locally) in the projected space. Each such a margin is set to the difference between the average distance to heterogeneous neighbors and the average distance to the homogeneous neighbors. [5] adopted a similar strategy that is based on the use of similar and dissimilar samples. Maximally Collapsing Metric Learning (MCML) algorithm [6] generates a metric (from which a linear

M. Kamel and A. Campilho (Eds.): ICIAR 2013, LNCS 7950, pp. 116–123, 2013.

transform is estimated) by trying to map all samples in the same class to a single point and push samples in other classes infinitely far away.

Marginal Fisher Analysis (MFA) [7] is introduced to remedy some of the shortcomings of the FDA technique. It is intended to perform local discrimination between classes. Whenever the training data set is small, MFA cannot directly be used with the original high-dimensional samples. This is referred to as the small sample size (SSS) phenomenon that happens whenever the feature dimension is higher than the number of examples. The classic remedy was using the projection of the raw data (e.g., using (PCA)). [8], introduces Exponential Discriminant Analysis (EDA) approach that is based on the exponential of the global within-class and between class covariance matrices. The EDA approach overcomes the SSS problem but it still similar to FDA framework in the sense that it does not take into account the local structures of the data.

In this paper, we propose two regularization frameworks that solve the SSS problem associated with MFA. Our frameworks can retain the discriminant information discarded by using the PCA pre-stage in MFA. The remainder of the paper is organized as follows. Section 2 reviews the MFA method. Section 3 describes our proposed frameworks. Experimental results obtained with four face data sets are presented in Section 4.

2 Review of Marginal Fisher Analysis (MFA)

The goal of MFA is to compute a transform that maximizes the distance between heterogeneous data samples and makes the data samples belonging to the same class closer to each other. We assume that we have a set of N labeled examples $\{\mathbf{x}_i\}_{i=1}^{N} \subset \mathbb{R}^D$. In order to find the discriminant structure of the data manifold, two graphs will be reconstructed: the within-class graph G_w (intrinsic graph) and between-class graph G_b (penalty graph). Let $l(\mathbf{x}_i)$ be the class label of $\mathbf{x}_i$. For each data sample $\mathbf{x}_i$, two subsets, $N_w(\mathbf{x}_i)$ and $N_b(\mathbf{x}_i)$ are computed. $N_w(\mathbf{x}_i)$ contains the neighbors sharing the same label with $\mathbf{x}_i$, while $N_b(\mathbf{x}_i)$ contains the neighbors having different labels. Those two sets are usually estimated using two nearest neighbor graphs: one graph is constructed for the data samples having the same label (this graph will have a parameter denoted by K_1). and one graph for the data samples with different label (this graph will have a parameter denoted by K_2). K_1 and K_2 can be selected empirically. Each of these graphs, G_w and G_b, is represented by its weight (affinity) matrix $\mathbf{W}_w$ and $\mathbf{W}_b$, respectively. The elements of these symmetric matrices are given by:

$$W_{w,ij} = \begin{cases} sim(\mathbf{x}_i, \mathbf{x}_j) \text{ if } \mathbf{x}_j \in N_w(\mathbf{x}_i) \text{ or } \mathbf{x}_i \in N_w(\mathbf{x}_j) \\ 0, otherwise \end{cases} \tag{1}$$

$$W_{b,ij} = \begin{cases} 1 \text{ if } \mathbf{x}_j \in N_b(\mathbf{x}_i) \text{ or } \mathbf{x}_i \in N_b(\mathbf{x}_j) \\ 0, otherwise \end{cases} \tag{2}$$

where $sim(\mathbf{x}_i, \mathbf{x}_k)$ encodes the similarity between sample $\mathbf{x}_i$ and sample $\mathbf{x}_k$. This function can be set to the Kernel heat or the cosine.

Any linear embedding method aims at computing a matrix transform that projects $\mathbf{x}_i$ into $\mathbf{A}^T\,\mathbf{x}_i$ (low dimensional representation of $\mathbf{x}_i$). MFA estimates the unknown, $\mathbf{A}$, that simultaneously maximizes the margins between heterogenous samples and moves the homogeneous samples closer to each other (after the transformation). Mathematically, this leads to:

$$\min_{\mathbf{A}} \frac{1}{2} \sum_{i,j} \|\mathbf{A}^T(\mathbf{x}_i - \mathbf{x}_j)\|^2 W_{w,ij} = \min_{\mathbf{A}} tr\left(\mathbf{A}^T\,\mathbf{X}\,\mathbf{L}_w\,\mathbf{X}^T\,\mathbf{A}\right) \tag{3}$$

$$\max_{\mathbf{A}} \frac{1}{2} \sum_{i,j} \|\mathbf{A}^T(\mathbf{x}_i - \mathbf{x}_j)\|^2 W_{b,ij} = \max_{\mathbf{A}} tr\left(\mathbf{A}^T\,\mathbf{X}\,\mathbf{L}_b\,\mathbf{X}^T\,\mathbf{A}\right) \tag{4}$$

where $tr(\cdot)$ denotes the matrix trace operator, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ is the data matrix, $\mathbf{L}_w = \mathbf{D}_w - \mathbf{W}_w$ is the Laplacian matrix of the graph G_w, $\mathbf{D}_w$ is the diagonal weight matrix, whose diagonal elements are column (or row, since $\mathbf{W}_w$ is symmetric) sums of $\mathbf{W}_w$.

The two criteria, Eq. (3) and Eq. (4), can be merged into one criterion that should be maximized:

$$J = \frac{tr\left(\mathbf{A}^T\,\mathbf{X}\,\mathbf{L}_b\,\mathbf{X}^T\,\mathbf{A}\right)}{tr\left(\mathbf{A}^T\,\mathbf{X}\,\mathbf{L}_w\,\mathbf{X}^T\,\mathbf{A}\right)} = \frac{tr\left(\mathbf{A}^T\,\widetilde{\mathbf{S}}_b\,\mathbf{A}\right)}{tr\left(\mathbf{A}^T\,\widetilde{\mathbf{S}}_w\,\mathbf{A}\right)} \tag{5}$$

where the symmetric matrix $\widetilde{\mathbf{S}}_b = \mathbf{X}\,\mathbf{L}_b\,\mathbf{X}^T$ is the locality preserving between class scatter matrix, and the symmetric matrix $\widetilde{\mathbf{S}}_w = \mathbf{X}\,\mathbf{L}_w\,\mathbf{X}^T$ is the locality preserving within class scatter matrix. Maximizing the trace ratio (5) can be replaced by the simpler form:

$$\max_{\mathbf{A}} tr\left\{\left(\mathbf{A}^T\,\widetilde{\mathbf{S}}_w\,\mathbf{A}\right)^{-1}\left(\mathbf{A}^T\,\widetilde{\mathbf{S}}_b\,\mathbf{A}\right)\right\} \tag{6}$$

The columns of the unknown transform $\mathbf{A}$ will be obtained by the generalized eigenvectors associated with the largest eigenvalues of:

$$\widetilde{\mathbf{S}}_b\,\mathbf{a} = \lambda\,\widetilde{\mathbf{S}}_w\,\mathbf{a} \tag{7}$$

The Small Sample Size problem. In many practical cases such as face recognition, both matrices $\mathbf{X}\mathbf{L}_b\mathbf{X}^T$ and $\mathbf{X}\mathbf{L}_w\mathbf{X}^T$ can be rank deficient. Indeed, very often the number of the training samples, N, is much smaller than the image dimension, D. This is referred to as the Small Sample Size (SSS) problem. In order to avoid getting singular matrices, the classical way is to project original high-dimensional data onto a PCA subspace so that the resulting matrices $\mathbf{X}\mathbf{L}_b\mathbf{X}^T$ and $\mathbf{X}\mathbf{L}_w\mathbf{X}^T$ are non-singular.

3 Proposed Schemes for Overcoming the SSS Problem

The SSS problem associated with MFA was solved by applying a PCA on the high-dimensional data. This process removes the null spaces of $\mathbf{X}\mathbf{L}_b\mathbf{X}^T$ and

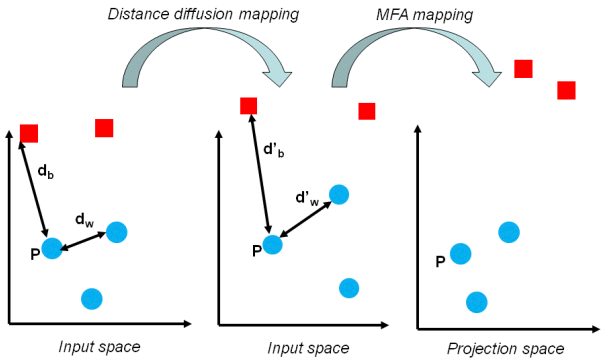

Fig. 1. The expected embedding carried out by EMFA

$\mathbf{X}\mathbf{L}_w\mathbf{X}^T$. Thus, by adopting PCA as an pre-stage in MFA framework some discriminative knowledge can be lost and will not exploited by the MFA framework. In this section, we propose two regularization schemes. The first one is based on regularizing the within-class scatter matrix. The second one is based on the use of Exponential matrices.

Regularized MFA (RMFA). Whenever the within-class matrix $\widetilde{\mathbf{S}}_w$ is singular, solving (7) directly will not be feasible. Therefore, the idea is to remove the singularity of $\widetilde{\mathbf{S}}_w$ by adding a regularization term. Therefore, the regularized version of MFA (RMFA) consists in estimating the generalized eigenvectors given by:

$$\widetilde{\mathbf{S}}_b\,\mathbf{a} = \lambda\,(\widetilde{\mathbf{S}}_w + \beta\,tr(\widetilde{\mathbf{S}}_w)\,\mathbf{I})\,\mathbf{a} \tag{8}$$

where β is a positive scalar and $\mathbf{I}$ is the $D \times D$ identity matrix. This regularization is linked to ridge regression in which the L_2 norm of the unknown transform [9] is minimized.

Exponential MFA (EMFA). The exponential of an $N \times N$ matrix $\mathbf{F}$ is given by [8]:

$$\exp(\mathbf{F}) = \mathbf{I} + \mathbf{F} + \frac{\mathbf{F}^2}{2!} + \ldots + \frac{\mathbf{F}^m}{m!} + \ldots$$

where $\mathbf{I}$ is the $N \times N$ identity matrix. Matrix exponential has the following interesting property:

Property 1. If $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N$ are eigenvectors of $\mathbf{F}$ that are associated to the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$, then $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N$ are also eigenvectors of $\exp(\mathbf{F})$ that are associated with the eigenvalues $e^{\lambda_1}, e^{\lambda_2}, \ldots, e^{\lambda_N}$. It is well known that the obtained matrix is non-singular.

The exponential version of MFA (EMFA) is got by inserting the exponential of the matrices $\widetilde{\mathbf{S}}_b$ and $\widetilde{\mathbf{S}}_w$ into the framework of MFA. As a result of this replacement, two beneficial effects on the whole embedding will be obtained: (i) the SSS problem will be overcome, and (ii) a distance diffusion mapping will be applied. These effects are resulting from property 1 (more explanation can be found in [8]). The second effect is implicit and has similar properties of the kernel methods used to get non-linear version of classic linear embedding methods such as Kernel PCA and Kernel FDA. The only difference is that EMFA works on the scatter matrices, while the kernel methods work on the original variables.

Figure 1 depicts a geometrical representation of the two processes that are induced by the EMFA method. The novel score to be optimized will be given by:

$$\max_{\mathbf{A}} tr\left\{\left(\mathbf{A}^T \exp(\widetilde{\mathbf{S}}_w)\,\mathbf{A}\right)^{-1}\left(\mathbf{A}^T \exp(\widetilde{\mathbf{S}}_b)\,\mathbf{A}\right)\right\} \tag{9}$$

The unknown $\mathbf{A}$ is given by the generalized eigenvectors of the following:

$$\exp(\widetilde{\mathbf{S}}_b)\,\mathbf{a} = \lambda\,\exp(\widetilde{\mathbf{S}}_w)\,\mathbf{a} \tag{10}$$

Note that $\widetilde{\mathbf{S}}_b$ and $\widetilde{\mathbf{S}}_w$ should be normalized, because $\exp(\widetilde{\mathbf{S}}_b)$ and $\exp(\widetilde{\mathbf{S}}_w)$ may have large numbers. We use Frobenius norm in order to normalize these matrices. However, this normalization may deteriorate the diffusion distance property induced by the use of matrix exponential. For this reason, we add two scaling parameters σ_b and σ_w that re-scale the normalized matrices $\widetilde{\mathbf{S}}_b$ and $\widetilde{\mathbf{S}}_w$, respectively. Finding the best values of these two parameters is carried out using the Differential Evolution algorithm [10] that maximizes the recognition rate over a validation set.

4 Performance Study

Databases. To verify the effectiveness of our proposed frameworks, we applied them to the face recognition problem. Four public face data sets are considered. Some images from PIE and FERET databases are illustrated in Figure 2. In our experiments, all face images are resized to 32×32.

1. **Yale**[1]**:** The YALE face database contains 15 persons. Each person has 11 images. This database shows variations in facial expression and in lighting.
2. **PIE**[2]**:** Our experiments use a subset containing 1926 images of 68 individuals. The images contain variations related to poses, illumination, and facial expression.
3. **PF01**[3]**:** It contains 103 persons. Each person has 17 images (1 normal face, 4 illumination modes, 8 pose modes, 4 expression modes) per individual.

[1] `http://see.xidian.edu.cn/vipsl/database_Face.html`
[2] `http://www.ri.cmu.edu/projects/project_418.html`
[3] `http://nova.postech.ac.kr/special/imdb/imdb.html`

4. **FERET**[4]**:** The proposed method is evaluated on a subset of FERET database, which includes 1400 images of 200 distinct subjects, each subject has seven images. The subset involves variations in facial expression, illumination and pose.

Tuning the scaling parameters. Before presenting the method comparison, we first show the usefulness of tuning the scaling parameters σ_b and σ_w for the EMFA scheme. Table 1 illustrates the recognition rate obtained over a validation subset of the PF01 dataset. The first column corresponds to the recognition rate obtained with the raw normalization of the locality preserving matrices (Frobenius normalization) for which σ_b and σ_w are both set to one. The remaining columns depicts the best solution (recognition rate) obtained by three successive iterations of the Differential Evolution algorithm. We observe that in general the DE algorithm has converged in only 2 iterations.

In another experiment, we consider the PIE dataset. We split it into three parts: training part (15 images per person), validation part (3 images per person) and test part (10 images per person). The training and validation parts are used for inferring the best linear transform as well as the best scaling parameters (σ_b and σ_w) using the Differential Evolution algorithm. Once these parameters are estimated the recognition rate on the test part is estimated. Table 2 illustrates the test recognition rate for ten random splits of the PIE dataset. As can be seen, the parameter tuning was very useful for enhancing the distance diffusion mapping of the EMFA framework.

Fig. 2. Some images in FERET database (top) and in PIE database (bottom)

Table 1. Recognition rates on a given split of the PF01 dataset using the EMFA method. The first column corresponds to the recognition rate obtained with the raw normalization of the locality preserving matrices (Frobenius normalization) for which σ_b and σ_w are both set to one. The remaining column depicts the best solution obtained by three successive iterations of the DE algorithm.

DE iterations	Frobenius norm	Ite. 1	Ite. 2	Ite. 3
Recognition rate	72.12	80.99	83.17	83.17

[4] `http://www.itl.nist.gov/iad/humanid/feret/`

Table 2. PIE recognition rate using the EMFA method

Split	1	2	3	4	5	6	7	8	9	10	*Average*
No tuning	83.5	81.9	71.9	83.6	82.5	77.5	81.5	79.9	76.5	83.9	**80.0**
With tuning	90.5	88.6	81.2	91.6	88.3	87.5	92.0	91.6	86.3	90.2	**88.6**

Table 3. Best average recognition rate (%) over 10 random splits using some embedding methods as well as the proposed regularization schemes (see text)

Method	*Yale (3)*	*PIE (5)*	*PF01 (3)*	*FERET (3)*
PCA	86.0	35.2	43.2	61.3
FDA	81.9	62.9	60.5	67.3
EDA	89.3	65.1	63.3	69.9
MCML	88.6	55.7	55.1	69.3
MFA	88.4	60.5	58.6	66.3
RMFA	92.7	67.7	68.7	72.0
EMFA	**93.1**	**70.5**	**70.4**	**74.5**

Table 4. Best average recognition rate (%) (see text)

Method	*Yale (7)*	*PIE (15)*	*PF01 (7)*	*FERET (5)*
PCA	88.9	55.8	53.3	68.2
FDA	91.6	85.9	74.1	81.0
EDA	94.7	86.4	75.0	81.8
MCML	95.3	81.6	69.3	79.6
MFA	93.7	85.0	72.3	79.7
RMFA	96.8	87.7	**82.0**	84.6
EMFA	**97.2**	**89.3**	81.1	**86.6**

Experimental results. Each data set is randomly partitioned into ten training/testing splits. For every person, we randomly selected l images as training examples, and the remaining images were used as test images. From the learning samples, a face subspace is built through the estimation of a linear transform using the following approaches: PCA, FDA, EDA, MCML, MFA, and the proposed schemes RMFA and EMFA. The FDA and MFA methods that suffer from the SSS problem used a PCA projection that retained 95% of the total variability of the training data. For all mapping methods, a test image is projected using the estimated the associated linear transform. The recognition is then performed in the novel projected subspace using the Nearest Neighbor classifier. The process is repeated for all (train/test) splits. We calculate the average recognition rate over these ten splits. In general, the recognition rate depends on the retained dimension of the mapping. Therefore, the average recognition rate will be a curve giving the recognition as a function of this retained dimension. Table 3 illustrates the best average recognition rate (%) over 10 random splits using the PCA, FDA,

EDA, MCML, MFA, RMFA and EMFA methods. For RMFA and EMFA the results correspond to the best performance over the tuning parameters. The results were obtained with Yale, PIE, PF01, and FERET data sets with small training sets. Table 4 illustrates the same results of Table 3 but this time the number of training images per person was increased. We can observe that: (1) EDA is superior to MFA and FDA, (2) both proposed schemes RMFA and EMFA are superior to EDA and to the classic regularization (PCA followed by MFA) and, (3) for many cases the EMFA scheme outperformed the RMFA scheme.

5 Conclusion

We proposed two solution schemes for overcoming the SSS problem of the Marginal Fisher Analysis method. The first scheme uses ridge regression regularization. The second scheme uses matrix exponential and introduces an implicit distance diffusion mapping. It integrates a similar effect to the non-linear Kernel-based embedding. The experiments are conducted on four face data sets. We have shown that the proposed schemes gave better results than using the classical solution based a PCA pre-stage. We found that in general the second scheme gave more accurate results than the first regularization scheme based on ridge regression.

Acknowledgment. This work was supported by the Spanish Government under the project TIN2010-18856.

References

1. Li, X., Lin, S., Yan, S., Xu, D.: Discriminant locally linear embedding with high-order tensor data. IEEE Trans. Syst., Man, Cybern. B: Cybern 32, 342–352 (2008)
2. Li, H., Jiang, T., Zhang, K.: Efficient and robust feature extraction by maximum margin criterion. IEEE Trans. on Neural Networks 17, 157–165 (2006)
3. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, New York (1990)
4. Wang, F., Wang, X., Zhang, D., Zhang, C., Li, T.: Marginface: A novel face recognition method by average neighborhood margin maximization. Pattern Recognition 42, 2863–2875 (2009)
5. Alipanahi, B., Biggs, M., Ghodsi, A.: Distance metric learning vs. Fisher discriminant analysis. In: AAAI Conference on Artificial Intelligence (2008)
6. Globerson, A., Roweis, S.: Metric learning by collapsing classes. In: Conference on Advances in Neural Information Processing Systems (2006)
7. Yan, S., Xu, D., Zhang, B., Zhang, H.J.: Graph embedding: A general framework for dimensionality reduction. In: Int. Conference on Computer Vision and Pattern Recognition (2005)
8. Zhang, T., Fang, B., Tang, Y., Shang, Z., Xu, B.: Generalized discriminant analysis: A matrix exponential approach. IEEE Transactions on Systems, Man, and Cybernetics 40, 186–197 (2010)
9. Zhang, Z., Dai, G., Xu, C., Jordan, M.: Regularized discriminant analysis, ridge regression and beyond. Journal of Machine Learning Research 11, 2199–2228 (2010)
10. Price, K.V., Lampinen, J.A., Storn, R.M.: Differential Evolution: A Practical Approach To Global Optimization. Springer (2005)