The First International Symposium on Optimization and Systems Biology (OSB'07)
Beijing, China, August 8–10, 2007
, pp. 250–258

A Discriminate Multidimensional Mapping for Small Sample Database*

Xiaoqin Hu Ling Jing† Qian Li

College of Science, China Agricultural University, 100083, Beijing, P.R.China

Abstract In this paper, we develop a new feature extraction method, called discriminate multidimensional mapping (DMM), which is especially effective for small sample database. The algorithm takes advantages of MDS and LDA, meanwhile avoids their disadvantages. We give theoretical analysis of this algorithm, and demonstrate with examples that DMM is effective in practical use.

1 Introduction

In many classification problems, such as face recognition and gene expression analysis, the data set is small but the dimension of sample is high, so feature extraction is essential before classification. There are a number of known dimensionality reduction techniques: principal component analysis (PCA) and multidimensional scaling (MDS)[1] are linear methods; the examples of nonlinear method are Locally linear embedding (LLE)[2], isometric feature mapping (Isomap)[3], Laplacian eigenmap[4], self-organizing mapping (SOM)[5] and kernel principal component analysis (KPCA)[6].

The above are all unsupervised methods, which are effective in finding compact representations and useful for data interpolation and visualization. But they are suboptimal from classification viewpoint because the information carried by class labels is lost. On the other hand, supervised methods such as linear discriminant analysis (LDA)[7] have shown their successes in pattern classification.

But compared with unsupervised methods MDS, LDA is prone to overfitting when the training data set is small and the dimension is large, which is often the case in face recognition and gene expression analysis.

To take advantages of MDS and LDA, meanwhile to avoid their disadvantages, we develop a new method called discriminant multidimensional mapping(DMM). The rest of this paper is organized as follows. In section 2, classical MDS and LDA are introduced. Discriminant multidimensional mapping(DMM) is discussed in detail in section 3. Experimental results are presented in section 4, followed by conclusions in section 5.

*This work is supported by the National Natural Science Foundation of China (No. 10631070).
†Corresponding author. E-mail: jingling_aaa@163.com, Tel:+86-10-62736511

2 Introduction of classical MDS and LDA

2.1 Classical MDS [8][9]

Given suitable information (such as similarity or dissimilarity measure) about a collection of N objects, the task of MDS is to embed the objects as points in a low-dimensional Euclidean space , while preserving the geometry as faithfully as possible. If the dissimilarity measure is Euclidean distance, it is classical MDS.

Consider a set of N points $\{x_i\}_{i=1}^N$, $x_i \in R^m$. We can proceed as follows: without loss of generality, assume that the points are centered, i.e. $\overline{X} = \sum_{i=1}^{N} x_i = 0$. The squared pairwise Euclidean distance

$$d^2(x_i,x_j) = \| x_i - x_j \|_2^2 = x_i^T x_i - 2x_i^T x_j + x_j^T x_j \tag{1}$$

Let the N-dimensional vector $\psi = [x_1^T x_1, \cdots, x_N^T x_N]^T$. Then the squared-distance matrix $\triangle = [d^2(x_i,x_j)]_{i,j=1}^N$ can be written as

$$\triangle = \psi e^T - 2X^T X + e\psi^T \tag{2}$$

where $X = [x_1, \cdots, x_N]_{m\times N}$, and $J = I - ee^T/N$, $e = [1, \cdots, 1]^2$. Then it follows that

$$H \equiv -J\triangle J/2 = X^T X \tag{3}$$

Compute the eigenvalues of H together with an orthonormal set of eigenvectors. Write λ_i for the i-th largest positive eigenvalue and v_i for the corresponding eigenvector (written as a column vector).

$$H = U diag(\lambda_1, \cdots, \lambda_N) U^T \qquad \lambda_1 \geq \cdots \geq \lambda_N \tag{4}$$

Then the required p-dimensional embedding vectors $\{y_i\}_{i=1}^N$, $y_i \in R^p$ are given by the columns of the following matrix:

$$Y = [y_1, \cdots, y_N] = diag(\lambda_1^{1/2}, \cdots, \lambda_p^{1/2}) U^{*^T} \qquad U^* = [v_1, v_2, \cdots, v_p] \tag{5}$$

Classical MDS has many appealing features. When the given metric on the input data points truly has a low-dimensional Euclidean structure, classical MDS is guaranteed to find a Euclidean embedding which exactly preserves metric. The required storing space is $O(N^2)$ and the computational complexity is $O(N^3 + pN^2)$.

2.2 Linear Discriminant analysis (LDA)

LDA is a well-known supervised technique for dealing classification problems. It is a derivative of Fisher's Linear Discriminant (FLD) which maximizes the ratio of between-class scatter to that of within-class scatter.

Given a set of N points $\{x_i\}_{i=1}^N$, $x_i \in R^m$ and each point belongs to one of the c class $\{z_i\}_{i=1}^c$. The between-class and within-class scatter matrixes are defined as

$$S_B = \sum_{i=1}^{c} \frac{N_i}{N} \sum_{j=1}^{c} (\mu_i - \mu_j)(\mu_i - \mu_j)^T \tag{6}$$

$$S_W = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N_i}(x_j^i - \mu_i)(x_j^i - \mu_i)^T \tag{7}$$

where μ is the mean of all points, μ_i is the mean of class z_i, x_j^i denote the j-th sample in the i-th class.

Let us consider a linear transformation mapping the original m-dimensional space into a p-dimensional feature space, where $p < m$. The new feature vector $y_k \in R^p$ is defined by the following linear transformation:

$$y_k = W^T x_k \qquad k = 1,2,\cdots,N \tag{8}$$

where $W \in R^{m\times p}$ is a matrix with orthonormal columns. Then after applying the linear transformation W, the scatter of the $\{y_i\}_{i=1}^N$ is $W^T S_B W$ and $W^T S_W W$. The optimal projection W^* is chosen as:

$$J(W^*) = \underset{W}{argmax}\frac{\mid W^T S_B W \mid}{\mid W^T S_W W \mid} = [w_1^*, w_2^*, \cdots, w_p^*] \tag{9}$$

where $\{w_i^*\}_{i=1}^p$ is the set of generalized eigenvectors of $S_W^{-1}S_B$, corresponding to the p largest generalized eigenvalues $\{\lambda_i\}_{i=1}^p$. The rank of S_B is $c-1$ or less because it is the sum of c matrixes of rank one or less. Thus, there are at most $c-1$ nonzero eigenvectors [10].

The computational complexity and memory requirement of LDA are dominated by the calculation of $S_W^{-1}S_B$, so LDA requires $O(m^2)$ space and $O(Nm^2 + m^3)$ operations.

3 Discriminant Multidimensional Mapping (DMM)

LDA is developed based on classification principle, but is too expensive in practice for high-dimensional database as it's computation is determined by the number of dimension. Classical MDS' computation is determined by the number of samples, but it's object is to preserve the distances of the embedded points as faithfully as possible which is less appropriate for classification. The advantages and disadvantages of both lead us to think: we can first run classical MDS to embed the original points in a low-dimensional Euclidean space R^N, then map the points in R^N to a less-dimensional space by LDA.

Inspired by the opinion, we develop a new method called discriminant multidimensional mapping (DMM). Before introducing DMM, we first give two theorems.

Theorem 1: Given a set of N points which we wish to embed in the Euclidean space R^k. δ_j denotes the vector of squared distances from the point x_j to $x_i, i = 1,\cdots,N$, and $\delta_\mu = (\delta_1 + \delta_2 + \cdots + \delta_N)/N$. $L = [v_1^T/\sqrt{\lambda_1} \quad v_2^T/\sqrt{\lambda_2} \quad \cdots \quad v_m^T/\sqrt{\lambda_m}]^T$, where v_i, λ_i are defined as (4). Then the embedding vector y_j(5) get by classical MDS can also be given by the formula:

$$y_j = -\frac{1}{2}L(\delta_j - \delta_\mu) \tag{10}$$

Proof. Let $\triangle = (\delta_1, \cdots, \delta_N), J = I - ee^T/N$. Note that $\delta_j - \delta_\mu$ is the j-th column of $\triangle J$, thus what we have to prove is $-\frac{1}{2}L\triangle J = Y$.

Because $J^2 = J$, $Hv_i = \lambda_i v_i$ and $H \equiv -J\triangle J/2$, $HJv_i = \lambda_i v_i$. So we get $Jv_i = v_i$, this is $LJ = J$. Now we have to prove is $-\frac{1}{2}LJ\triangle J = Y$, this is $LH = Y$. For each eigenvector v_i^T of H we have

$$\frac{v_i^T H}{\sqrt{\lambda_i}} = \frac{\lambda_i v_i^T}{\sqrt{\lambda_i}} = \sqrt{\lambda_i} v_i^T \tag{11}$$

so the i-th row of LH equals the i-th Y. This completes the proof. □

Theorem 2: Let $X = [x_1, \cdots, x_N]$, $Y = [y_1, \cdots, y_N]$, and they have the relation $Y = W_1^T X$. applying LDA to X and Y respectively, get optimal projections W_2 and W_3. Then

$$W_2 = W_1 W_3 \tag{12}$$

Proof. The between-class and within-class scatter matrixes of X are S_B^1 and S_W^1. S_B^2 and S_W^2 are those of Y. Using (6) and (7), it is easy to know:

$$S_B^2 = W_1^T S_B^1 W_1 \qquad S_B^2 = W_1^T S_W^1 W_1 \tag{13}$$

W_2, W_3 are defined as follows:

$$W_2 = \underset{W}{argmax} \frac{\mid W^T S_B^1 W \mid}{\mid W^T S_W^1 W \mid} \qquad W_3 = \underset{W}{argmax} \frac{\mid W^T S_B^2 W \mid}{\mid W^T S_W^2 W \mid} \tag{14}$$

Using (13), W_3 can be written as:

$$W_3 = \underset{W}{argmax} \frac{\mid (W_1 W)^T S_B^1 (W_1 W) \mid}{\mid (W_1 W)^T S_W^1 (W_1 W) \mid} \tag{15}$$

Compare (14) and (15), it is obvious that $W_2 = W_1 W_3$. □

Based on the above theorems, we develop the DMM algorithm. Consider a data set $\{x_i\}_{i=1}^N$, $x_i \in R^m$ and $N \ll m$, the number of class is c. The embedding vectors $\{z_i\}_{i=1}^N$, $z_i \in R^{(c-1)}$ is we wish to get. Applying classical MDS to $\{x_i\}_{i=1}^N$, based on **theorem 1**, we first have:

$$y_i = -\frac{1}{2} L(\delta_i - \delta_\mu) \tag{16}$$

where δ_i and δ_μ are defined as in **theorem 1**. Let $-\frac{1}{2}L = W_1^T$, then $y_i = W_1^T(\delta_i - \delta_\mu)$, $y_i \in R^N$. By applying LDA to $\{y_i\}_{i=1}^N$, we get the embedding vectors $z_i = W_2^T y_i, z_i \in R^{c-1}$, where W_2 is obtained by optimal objection(9). Concluding the above, we get

$$z_i = W_2^T W_1^T (\delta_i - \delta_\mu) \tag{17}$$

Let $W^* = W_1 W_2$

$$z_i = W^{*T}(\delta_i - \delta_\mu) \tag{18}$$

Base on **theorem 2** we can see

$$J(W^*) = \underset{W}{argmax} \frac{\mid W^T S_B^* W \mid}{\mid W^T S_W^* W \mid} \tag{19}$$

Where S_B^* and S_W^* are the between-class and within-class scatter matrixes of $\{\delta_i - \delta_\mu\}_{i=1}^N$.

The DMM algorithm is summarized as follow:

1. Compute the squared distance matrix $\triangle = (\delta_1, \cdots, \delta_N)$, δ_i denotes the vector of squared distances from the i-th point to the others. Let $\delta_\mu = (\delta_1 + \delta_2 + \cdots + \delta_N)/N$。
2. Compute the within-class and between-class scatter matrixes S_W and S_B of $\{\delta_i - \delta_\mu\}_{i=1}^N$. Eigendecompositing $S_W^{-1} S_B$, we get the eigenvalue $\lambda_i (i = 1, \cdots, c-1)$ and the corresponding eigenvcetor v_i.
3. Define $W = [v_1, v_2, \cdots, v_{c-1}]$, then the embedding vector z_i can be obtained by the formula:
$$z_i = W^T(\delta_i - \delta_\mu) \tag{20}$$

4 Numerical experiments

Two classical pattern classification problems, face recognition and gene expression analysis, are considered in order to evaluate the performance of DMM's ability for small-sample and high-dimensional database.

4.1 Face recognition

In this subsection, we test several dimensional reduction methods using the publicly available AT&T and UMIST databases.

The AT&T (formerly Olivetti) face database contains 400 images of 40 people (http:// www.uk.research.att.com/facedatabase.html). They contain facial contours and vary in pose as well as scale. They original pixels are $112 \times 92 = 10304$, but for the complexity of Fisherface and PCA, we have to reduced to $23 \times 28 = 644$ for experiments. Figure 1. shows images of a few subjects.

The experiments are performed using the "leave-one-out" strategy. The training set are projected to a low-dimensional space and recognition is performed using a nearest neighbor classifier. The parameters, such as the number of neighbor in LLE and Isomap, the dimension of the embedding, are determined to achieve the lowest error rate by each method. For Fisherface and DMM, the points are projected automatically onto a $c-1$ space. The experimental results are shown in Table 1.

The UMIST Face Database consists of 575 images of 20 people (http://image s.ee.umist.ac.uk/danny/database.html). Each covering a range of poses from profile to frontal views. Subjects cover a range of race/sex/appearance. They original pixels are $112 \times 92 = 10304$, but for the same reason in AT&T experiment, we have to reduce to $23 \times 28 = 644$ for experiments. Some images of one people is published in

Figure 1: Face images in the AT&T database

Table 1: Results with the AT&T database

method	reduce space	erro rate(%)
eigenface	40	1.75(7/400)
fisherface	39	2.75(11/400)
MDS	28	2.75(11/400)
Isomap(k=110)	30	2.0(8/400)
LLE(k=40)	70	1.75(7/400)
DMM	39	0.5(2/400)

Figure 2. The experiments are performed in the same way as in AT&T. The experimental results are shown in Table 1. Figure 3. shows the samples of UMIST database projected onto the first two eigenvectors by DMM.

4.2 Gene expression data classification

Array technologies have made it possible to simultaneously monitor expression patterns of all genes in genome. The challenge now is to make sense of of such massive data sets. As the gene expression data are tens of thousands dimension in terms of a small number samples, usually those traditional methods are less appropriate for classifying this data.

Here we compared DMM with nonnegative matrix factorization (NMF) [12], a well known method in patterns of gene expression, to classify three cancer data sets. The experiments are performed using the "leave-one-out" strategy. DMM projects training set to a low-dimensional space and recognition is performed using a nearest

Figure 2: Face images in UMIST database

Table 2: Results with the UMIST database.

method	reduce space	erro rate(%)
eigenface	19	0.87(5/575)
fisherface	19	0.87(5/575)
MDS	40	0.70(4/575)
Isomap(k=70)	19	1.57(9/575)
LLE(k=40)	40	1.04(6/575)
DMM	19	0.35(2/575)

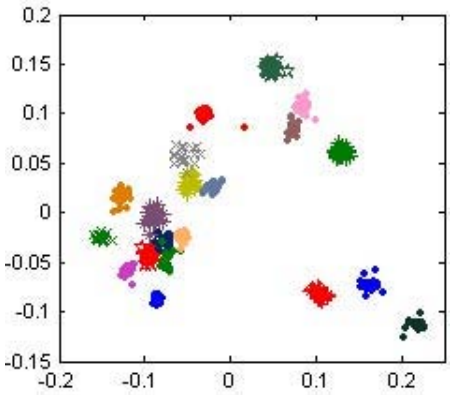

Figure 3: UMIST database projected onto the first two eigenvectors

neighbor classifier. NMF factors the training set to get the metagenes and testing set are classified by the most highly expressed metagene.

The first is leukemia data set with 5000 genes , which contains 38 samples: 27 acute myelogenous(AML) and 11 acute lymphoblastic leukemia(ALL). This data set has become a benchmark in the cancer classification community. It contains two ALL samples that are consistently misclassified by most methods. The second experiment is on a data set of 25 classic and 9 desmoplastic medulloblastoma tumors with 5893 genes. The third contains 7129 genes and 72 samples, and 25 of them are acute lymphoid leukemia (ALL), 47 of them are acute myeloid leukemia (AML). The error rates are shown in Table 3.

Table 3: Results on three gene data sets

data set	Leukemia	Medulloblastoma	MIT
NMF	7.9%(3/38)	38.24%(13/34)	26.39%(19/72)
DMM	2.63%(1/38)	20.59%(7/34)	1.39%(1/72)

5 Conclusion

Base on classical MDS and LDA, we develop a new feature extraction, called discriminate multidimensional mapping (DMM), which is especially effective for small sample database from the classification viewpoint. Compared with other dimensionality reduction techniques, DMM has several appealing features:

1. DMM do not need to select any parameters, while other methods perform drastically different with the parameters vary.
2. In every experiments, MDD performs better than other methods even their parameters are selected by achieving the lowest error rates.
3. The computational complexity of MDD is determined by the number of points. It is effective for small sample database in high-dimensional space.
4. It has a formula to embed a new sample to low-dimensional space, so it is convenient for out-of-sample.

Our future work will focus on the follows: we will extend DMM for high-dimensional and large-sample database by applying landmark, and generalize DMM to non-linear manifolds in the similarly way as Isomap.

References

[1] W.S. Torgerson. Multidimensional Scaling: Theory and Method. Psychometrika 1952, 17, 401-419.

[2] S.T. Roweis and L.K. Saul. Nonliear dimensionality reduction by locally linear cmbedding. Science, 2000, 290:2323-2326.

[3] J.B. Tenenbaum, V.(de) Silva , and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500):2319-2323, December (2000)

[4] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. Advances in Neural Information Processing Systems. 14, 2002.

[5] T. Kohonen. Self-organizing Maps. SPRINGER, Heidelberg, 2nd edition(1995).

[6] B. Scholkopf, A. Smolla and K. Muller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Computation, 10(5),1998.

[7] P. N. Belhumeur, J. P. Hespanha, D.J. Kriengman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. IEEE Trans. Pattern Anal. Mzch. Intell. 19(7), 1997, 711-720.

[8] H. Y. Zha and Z. Y. Zhang. Isometric Embedding and Contiunnm ISOMAP. Proceedings of the Twentieth International Conference on Mzchine Learning(ICML-2003), Waxhington DC,2003

[9] V. (de) Silva and J.B.Tenenbaum. Sparse multidimensional scaling using landmark points. (2004) available at http://pages.pomona.edu/vds04747/public/publications.html

[10] R. O. Duda, P. E. Hart, and R. L. Rivest. Pattern Classification. New York: Wiley-Interscience.

[11] M. H. Yang. Extended Isomap for Pattern Classification.

[12] J. P. Brunet, P. Tamayo, T. R. Golub, et al. Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the National Academy of Sciences, 2004, 101:4164-4169.