3D6-1

Dimensionality Reduction of Partially Labeled Multimodal Data

Yasushi Kitamura Masashi Sugiyama

Department of Computer Science, Tokyo Institute of Technology

Dimensionality reduction is one of the important preprocessing steps in the high-dimensional data analysis. In the semi-supervised learning scenario with partially labeled samples, we expect that samples in the same cluster are likely to share the common label (i.e. 'cluster assumption'), and based on this belief we classify the unlabeled samples. Therefore, when reducing the dimensionality of partially labeled samples, it is desirable to preserve the cluster structures of the data in addition to separating the labeled samples in different classes from each other. In this paper, we propose a new semi-supervised dimensionality reduction method that can achieve locality preservation and between-class separation at the same time. The proposed method has an analytic optimal solution and is computationally efficient. Simulations with benchmark data sets underline the usefulness of the proposed method.

1. Introduction

The goal of dimensionality reduction is to embed high-dimensional data samples in a low-dimensional space such that most of 'intrinsic information' contained in the original data is preserved. Once dimensionality reduction is carried out appropriately, the compact representation of the data can be used for various succeeding tasks such as visualization, classification, etc.

In the supervised learning scenario where data samples are accompanied with class labels, *Fisher discriminant analysis* (FDA) (Fisher 36) (Fukunaga 90) is a popular dimensionality reduction method. FDA seeks an embedding transformation such that the between-class scatter is maximized and the within-class scatter is minimized. FDA works very well if samples in each class are Gaussian with common covariance structure. However, it tends to give undesired results if samples in one class form several separate clusters (i.e., *multimodal*). To improve the performance of FDA, *local Fisher discriminant analysis* (LFDA) (Sugiyama 06) was proposed.

Although LFDA overcome the weakness of FDA, its performance tends to be degraded when only a small number of labeled samples are available. In such cases, it is effective to make use of *unlabeled* samples which are often available abundantly (i.e., semi-supervised learning). A requirement for the success of semi-supervised learning is the *cluster assumption*; samples in the same cluster have the same label (Chapelle 06). This implies that unsupervised dimensionality reduction methods which can preserve cluster structures would be useful in semi-supervised learning problems.

Locality-preserving projection (LPP) (He 04) is an unsupervised dimensionality reduction method that meets this requirement. LPP seeks an embedding transformation such that nearby sample pairs in the original high-dimensional space are kept close in the embedding space. Thus, LPP allows us to reduce the dimensionality of the data without losing the local structure.

In this paper, we propose a new semi-supervised dimensionality reduction method. Our approach is to combine LFDA and LPP in order to trade between-class separation with locality preservation. The proposed method includes FDA, LFDA, and LPP as special cases. Furthermore, the proposed method inherits the computational advantage of FDA, LFDA, and LPP, i.e., an analytic solution is available and can be computed based on the eigendecomposition. Thus the proposed method is computationally efficient and reliable. The usefulness of the proposed method is illustrated by experiments.

Contact: Yasushi Kitamura, 2-12-1-W8-74, O-okayama, Meguro-ku, Tokyo, 152-8552, Japan.
kitamura@sg.cs.titech.ac.jp

2. Linear Dimensionality Reduction

In this section, we formulate the problem of linear dimensionality reduction and review existing methods.

2.1 Formulation and Notation

Let $\boldsymbol{x}_i \in \mathbb{R}^d$ $(i = 1, 2, \ldots, n)$ be d-dimensional samples, and let $\boldsymbol{X}$ be the matrix of all samples:

$$\boldsymbol{X} \equiv (\boldsymbol{x}_1 | \boldsymbol{x}_2 | \cdots | \boldsymbol{x}_n). \tag{1}$$

Let $\boldsymbol{z} \in \mathbb{R}^r$ $(1 \leq r \leq d)$ be a low-dimensional representation of high-dimensional sample $\boldsymbol{x} \in \mathbb{R}^d$, where r is the dimensionality of the reduced space. Effectively we consider d to be large and r to be small, but not limited to such cases.

For the moment, we focus on linear dimensionality reduction, i.e., using a $d \times r$ transformation matrix $\boldsymbol{T}$, an embedded representation $\boldsymbol{z}$ of a sample $\boldsymbol{x}$ is given by

$$\boldsymbol{z} = \boldsymbol{T}^\top \boldsymbol{x}, \tag{2}$$

where $^\top$ denotes the transpose of a matrix or a vector.

When discussing supervised learning problems, we assume class labels $y_i \in \{1, 2, \ldots, c\}$ associated with the samples $\boldsymbol{x}_i$ are available, where c is the number of classes. We denote the number of samples in class $\ell \in \{1, 2, \ldots, c\}$ by n_ℓ.

Many dimensionality reduction techniques developed so far are based on the optimization problem of the following form.

$$\boldsymbol{T}_{OPT} \equiv \underset{\boldsymbol{T} \in \mathbb{R}^{d \times r}}{\operatorname{argmax}} \left[\operatorname{tr} \left(\left(\boldsymbol{T}^\top \underline{\boldsymbol{C}} \boldsymbol{T} \right)^{-1} \boldsymbol{T}^\top \overline{\boldsymbol{C}} \boldsymbol{T} \right) \right]. \tag{3}$$

Let $\{\boldsymbol{\varphi}_k\}_{k=1}^d$ be the eigenvectors associated with the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ of the following eigenvalue problem:

$$\overline{\boldsymbol{C}}\boldsymbol{\varphi} = \lambda \underline{\boldsymbol{C}}\boldsymbol{\varphi}. \quad (4)$$

Then a solution $\boldsymbol{T}_{OPT}$ is analytically given as follows (e.g., (Fukunaga 90))

$$\boldsymbol{T}_{OPT} = (\boldsymbol{\varphi}_1|\boldsymbol{\varphi}_2|\cdots|\boldsymbol{\varphi}_r). \quad (5)$$

2.2 Fisher Discriminant Analysis for Dimensionality Reduction

A popular supervised dimensionality reduction technique is *Fisher discriminant analysis* (FDA) (Fisher 36) (Fukunaga 90). Let $\boldsymbol{S}^{(w)}$ and $\boldsymbol{S}^{(b)}$ be the *within-class scatter matrix* and the *between-class scatter matrix*:

$$\boldsymbol{S}^{(w)} \equiv \sum_{\ell=1}^{c} \sum_{i:y_i=\ell} (\boldsymbol{x}_i - \boldsymbol{\mu}_\ell)(\boldsymbol{x}_i - \boldsymbol{\mu}_\ell)^\top, \quad (6)$$

$$\boldsymbol{S}^{(b)} \equiv \sum_{\ell=1}^{c} n_\ell (\boldsymbol{\mu}_\ell - \boldsymbol{\mu})(\boldsymbol{\mu}_\ell - \boldsymbol{\mu})^\top, \quad (7)$$

where $\boldsymbol{\mu}_\ell \equiv \sum_{i:y_i=\ell} \boldsymbol{x}_i/n_\ell$ and $\boldsymbol{\mu} \equiv \sum_{i=1}^n \boldsymbol{x}_i/n$. The FDA transformation matrix $\boldsymbol{T}_{FDA}$ is defined as

$$\boldsymbol{T}_{FDA} \equiv \underset{\boldsymbol{T}\in\mathbb{R}^{d\times r}}{\operatorname{argmax}} \left[\operatorname{tr}\left(\left(\boldsymbol{T}^\top \boldsymbol{S}^{(w)} \boldsymbol{T} \right)^{-1} \boldsymbol{T}^\top \boldsymbol{S}^{(b)} \boldsymbol{T} \right) \right]. \quad (8)$$

That is, FDA seeks a transformation matrix $\boldsymbol{T}$ such that the between-class scatter is maximized and the within-class scatter is minimized in the embedding space $\mathbb{R}^r$. A solution $\boldsymbol{T}_{FDA}$ is given by Eqs.(4) and (5) with $\overline{\boldsymbol{C}} = \boldsymbol{S}^{(b)}$ and $\underline{\boldsymbol{C}} = \boldsymbol{S}^{(w)}$.

The between-class scatter matrix $\boldsymbol{S}^{(b)}$ has at most rank $c-1$ (Fukunaga 90). This implies that FDA can find at most $c-1$ meaningful features; the remaining features found by FDA can be arbitrarily rotated in the null space of $\boldsymbol{S}^{(b)}$. This is an essential limitation of FDA for dimensionality reduction and is very restrictive in practice.

2.3 Locality-Preserving Projection

A useful unsupervised dimensionality reduction technique is *locality-preserving projection* (LPP) (He 04). Let $\boldsymbol{A}$ be the *affinity matrix*, i.e., the n-dimensional square matrix with $A_{i,j}$ being the affinity between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. We assume that $A_{i,j} \in [0,1]$; $A_{i,j}$ is large if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are 'close' and $A_{i,,j}$ is small if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are 'far apart'. There are several different manners of defining $\boldsymbol{A}$. All through this paper, we determine the affinity matrix $\boldsymbol{A}$ by the *local scaling method* (Zelnik-Manor 05). Let $\boldsymbol{S}^{(n)}$ and $\boldsymbol{S}^{(l)}$ be the *normalization matrix* and the *local scatter matrix* defined by

$$\boldsymbol{S}^{(n)} \equiv \boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^\top, \quad (9)$$

$$\boldsymbol{S}^{(l)} \equiv -\frac{1}{2} \sum_{i,j=1}^{n} W_{i,j} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top, \quad (10)$$

where $W_{i,j} \equiv A_{i,j}/n$ and $\boldsymbol{D}$ is the n-dimensional diagonal matrix with $D_{i,i} \equiv \sum_{j=1}^n W_{i,j}$. The LPP transformation matrix $\boldsymbol{T}_{LPP}$ is defined as

$$\boldsymbol{T}_{LPP} \equiv \underset{\boldsymbol{T}\in\mathbb{R}^{d\times r}}{\operatorname{argmax}} \left[\operatorname{tr}\left(\left(\boldsymbol{T}^\top \boldsymbol{S}^{(n)} \boldsymbol{T} \right)^{-1} \boldsymbol{T}^\top \boldsymbol{S}^{(l)} \boldsymbol{T} \right) \right]. \quad (11)$$

That is, LPP seeks a transformation matrix $\boldsymbol{T}$ such that *nearby* data pairs in the original space $\mathbb{R}^d$ are kept close in the embedding space $\mathbb{R}^r$. Thus, LPP tends to preserve the local structure of the data. $(\boldsymbol{T}^\top \boldsymbol{S}^{(n)} \boldsymbol{T})^{-1}$ works as a constraint to avoid degeneracy. A solution $\boldsymbol{T}_{LPP}$ is given by Eqs.(4) and (5) with $\overline{\boldsymbol{C}} = \boldsymbol{S}^{(l)}$ and $\underline{\boldsymbol{C}} = \boldsymbol{S}^{(n)}$.

2.4 Local Fisher Discriminant Analysis

FDA tends to perform poorly when the data has multimodality or outliers (Fukunaga 90). To cope with this problem, a localized variant of FDA called *local Fisher discriminant analysis* (LFDA) was proposed (Sugiyama 06). $\boldsymbol{S}^{(w)}$ and $\boldsymbol{S}^{(b)}$ defined by Eqs.(6) and (7) are expressed as

$$\boldsymbol{S}^{(w)} \equiv \frac{1}{2} \sum_{i,j=1}^{n} W_{i,j}^{(w)} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top, \quad (12)$$

$$\boldsymbol{S}^{(b)} \equiv \frac{1}{2} \sum_{i,j=1}^{n} W_{i,j}^{(b)} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top, \quad (13)$$

where

$$W_{i,j}^{(w)} \equiv \begin{cases} 1/n_\ell & \text{if } y_i = y_j = \ell, \\ 0 & \text{if } y_i \neq y_j, \end{cases} \quad (14)$$

$$W_{i,j}^{(b)} \equiv \begin{cases} 1/n - 1/n_\ell & \text{if } y_i = y_j = \ell, \\ 1/n & \text{if } y_i \neq y_j. \end{cases} \quad (15)$$

$1/n - 1/n_\ell$ in Eq.(15) is negative while $1/n_\ell$ and $1/n$ in Eqs.(14) and (15) are positive. This implies that if the data pairs in the same class are made close, the within-class scatter matrix $\boldsymbol{S}^{(w)}$ gets 'small' and the between-class scatter matrix $\boldsymbol{S}^{(b)}$ gets 'large'. On the other hand, if the data pairs in different classes are made further apart, the between-class scatter matrix $\boldsymbol{S}^{(b)}$ gets 'small'. Therefore, we may interpret FDA as keeping the sample pairs in the same class close and making the sample pairs in different classes apart.

Based on the above pairwise expression, the *local* within-class scatter matrix $\boldsymbol{S}^{(lw)}$ and the *local* between-class scatter matrix $\boldsymbol{S}^{(lb)}$ are defined as

$$\boldsymbol{S}^{(lw)} \equiv \frac{1}{2} \sum_{i,j=1}^{n} W_{i,j}^{(lw)} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top, \quad (16)$$

$$\boldsymbol{S}^{(lb)} \equiv \frac{1}{2} \sum_{i,j=1}^{n} W_{i,j}^{(lb)} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top, \quad (17)$$

where

$$W_{i,j}^{(lw)} \equiv \begin{cases} A_{i,j}/n_\ell & \text{if } y_i = y_j = \ell, \\ 0 & \text{if } y_i \neq y_j, \end{cases} \quad (18)$$

$$W_{i,j}^{(lb)} \equiv \begin{cases} A_{i,j}(1/n - 1/n_\ell) & \text{if } y_i = y_j = \ell, \\ 1/n & \text{if } y_i \neq y_j. \end{cases} \quad (19)$$

Namely, the values for the sample pairs in the same class are weighted according to the affinity $A_{i,j}$. This means that far apart sample pairs in the same class have less influence on $\boldsymbol{S}^{(lw)}$ and $\boldsymbol{S}^{(lb)}$. When $A_{i,j} = 1$ for all i, j, $\boldsymbol{S}^{(lw)}$ and $\boldsymbol{S}^{(lb)}$ are reduced to the original $\boldsymbol{S}^{(w)}$ and $\boldsymbol{S}^{(b)}$, respectively.

The LFDA transformation matrix $\boldsymbol{T}_{LFDA}$ is defined as

$$\boldsymbol{T}_{LFDA} \equiv \underset{\boldsymbol{T}\in\mathbb{R}^{d\times r}}{\operatorname{argmax}} \left[\operatorname{tr}\left(\left(\boldsymbol{T}^\top \boldsymbol{S}^{(lw)} \boldsymbol{T} \right)^{-1} \boldsymbol{T}^\top \boldsymbol{S}^{(lb)} \boldsymbol{T} \right) \right]. \tag{20}$$

That is, LFDA seeks a transformation matrix $\boldsymbol{T}$ such that nearby data pairs in the same class are made close and the data pairs in different classes are made apart; far apart data pairs in the same class are not imposed to be close. A solution $\boldsymbol{T}_{LFDA}$ is given by Eqs.(4) and (5) with $\overline{\boldsymbol{C}} = \boldsymbol{S}^{(lb)}$ and $\underline{\boldsymbol{C}} = \boldsymbol{S}^{(lw)}$. $\boldsymbol{S}^{(lb)}$ usually has full rank and thus LFDA can be applied to dimensionality reduction into *any* dimensionality.

3. Semi-Supervised Dimensionality Reduction

If only a small number of labeled samples are available, supervised dimensionality reduction methods tend to overfit the embedding space to the labeled samples; thus their performance can be heavily degraded. In such cases, it is effective to utilize unlabeled samples which are often available abundantly (i.e., semi-supervised learning). The success of semi-supervised learning is highly dependent on the *cluster assumption*; samples in the same cluster belong to the same class. This implies that cluster preserving dimensionality reduction methods such as LPP could be useful in semi-supervised learning problems.

Based on the above idea, in this section, we propose a new semi-supervised dimensionality reduction method by combining LFDA and LPP. From here on, we consider the case where some of the samples $\{\boldsymbol{x}_i\}_{i=1}^n$ are labeled and the rest are unlabeled.

3.1 Definition

The embedding transformations of LPP and LFDA can be analytically computed based on the eigendecompositions (See Sections 2.3 and 2.4). Our idea is to combine the eigenvalue problems of LPP and LFDA and solve them together. This allows us to maintain the computational efficiency and reliability of LPP and LFDA.

More specifically, for a weight $\beta \in [0,1]$, we solve the eigenvalue problem (4) with $\overline{\boldsymbol{C}} = \overline{\boldsymbol{S}}$ and $\underline{\boldsymbol{C}} = \underline{\boldsymbol{S}}$, where

$$\overline{\boldsymbol{S}} \equiv (1-\beta)\boldsymbol{S}^{(lb)} + \beta \boldsymbol{S}^{(l)}, \tag{21}$$

$$\underline{\boldsymbol{S}} \equiv (1-\beta)\boldsymbol{S}^{(lw)} + \beta \boldsymbol{S}^{(n)}. \tag{22}$$

Then a solution is still given by Eq.(5). Originally, LFDA is defined only for labeled samples. Therefore, when computing $\boldsymbol{S}^{(lb)}$ and $\boldsymbol{S}^{(lw)}$, we assign zero to $W_{i,j}^{(lb)}$ and $W_{i,j}^{(lw)}$ if at least one of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is unlabeled; if both of them are labeled, we compute $W_{i,j}^{(lb)}$ and $W_{i,j}^{(lw)}$ by Eqs.(19) and (18) as usual. Similarly, LPP is originally defined only for unlabeled samples. When computing $\boldsymbol{S}^{(l)}$ and $\boldsymbol{S}^{(n)}$, we treat all the samples $\{\boldsymbol{x}_i\}_{i=1}^n$ as unlabeled.

If $\beta = 0$, the proposed method is operated as fully supervised (i.e., all the unlabeled samples are discarded) and is reduced to LFDA. On the other hand, if $\beta = 1$, the proposed method is operated as fully unsupervised (i.e., all the labels are discarded) and is reduced to LPP. If $0 < \beta < 1$, the proposed method makes a compromise and gives an intermediate result between LFDA and LPP.

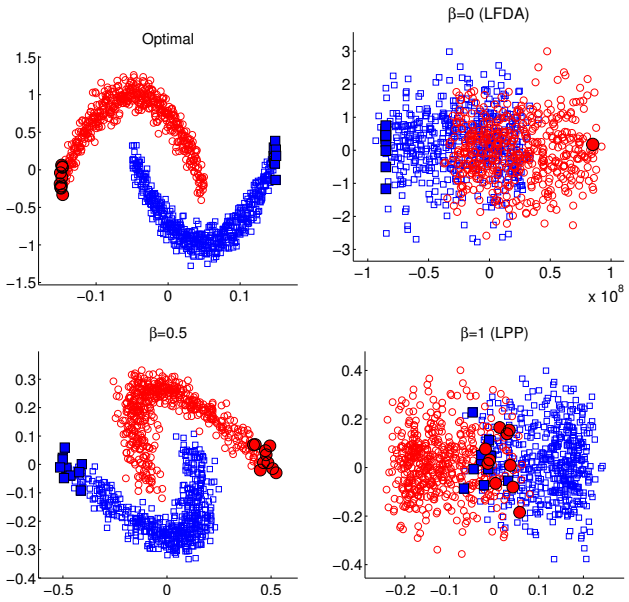

Figure 1: Embedded toy data samples by each method. Filled (unfilled) symbols are labeled (unlabeled) samples.

3.2 Illustrative Examples

In order to illustrate how the proposed method behave, we performed experiments with a 20-dimensional toy data set. The first two dimensions are 'two-moon' data and the others are Gaussian noise. We have 10 labeled samples and 490 unlabeled samples in each moon. The optimally embedded 2-dimensional samples are depicted in the top-left graph of Figure 1.

The samples embedded by LFDA is illustrated in the top-right graph of Figure 1. It shows that LFDA nicely separates the labeled samples in different classes from each other. However, the unlabeled samples (not used in LFDA) are mixed and hence the classification performance of unlabeled samples may be poor.

The samples embedded by LPP is illustrated in the bottom-right graph in Figure 1. It shows that two clusters are rather separated. However, since the label information is not used in LPP, the samples in different classes are mixed in the embedding space.

The samples embedded by the proposed method with $\beta = 0.5$ is illustrated in the bottom-left graph in Figure 1. Compared to LFDA and LPP, the cluster structures are rather preserved and labeled samples in different classes are well separable.

4. Semi-Supervised Classification Simulation

In this section, we evaluate the performance of the proposed method using the *IDA* data sets (Rätsch 01). As a performance measure, we adopt the misclassification rate

by a graph regularization method (Chapelle 06), i.e., using m labeled samples $\{\boldsymbol{x}_i, y_i\}_{i=1}^m$ $(1 \leq m < n)$ and $n-m$ unlabeled samples $\{\boldsymbol{x}_i\}_{i=m+1}^n$, predicted labels $\{\tilde{y}_i\}_{i=m+1}^n$ are obtained as

$$\underset{\{\tilde{y}_i\}_{i=1}^n}{\operatorname{argmin}} \left[\sum_{i=1}^{m} (\tilde{y}_i - y_i)^2 + \gamma \sum_{i,j=1}^{n} A_{i,j} (\tilde{y}_i - \tilde{y}_j)^2 \right], \qquad (23)$$

where we set $\gamma = 1$.

Figure 2 depicts the misclassification error rates as functions of the dimensionality of the embedding space. This illustrates an interesting tendency: When $\beta = 0$ (i.e., LFDA) works better than $\beta = 1$ (i.e., LPP), an intermediate β tends to outperform $\beta = 0$. On the other hand, when $\beta = 1$ works better than $\beta = 0$, an intermediate β tends to be outperformed by $\beta = 0$.

This tendency could be interpreted as follows. When $\beta = 0$ is better than $\beta = 1$, the cluster assumption may be unreliable and label information is essential. However, since only a small number of labeled samples are available, the embedding space found by $\beta = 0$ could be overfitted to the labeled samples. In such cases, the 'weak' cluster-preservation can improve the performance.

On the other hand, when $\beta = 1$ is better than $\beta = 0$, the cluster assumption would be highly reliable and local structure preservation is essential. In such cases, using label information can collapse local structures since (possibly nearby) labeled samples in different classes need to be separated from each other. As a result, $\beta < 1$ tends to yield an embedding space that is overfitted to the labeled samples.

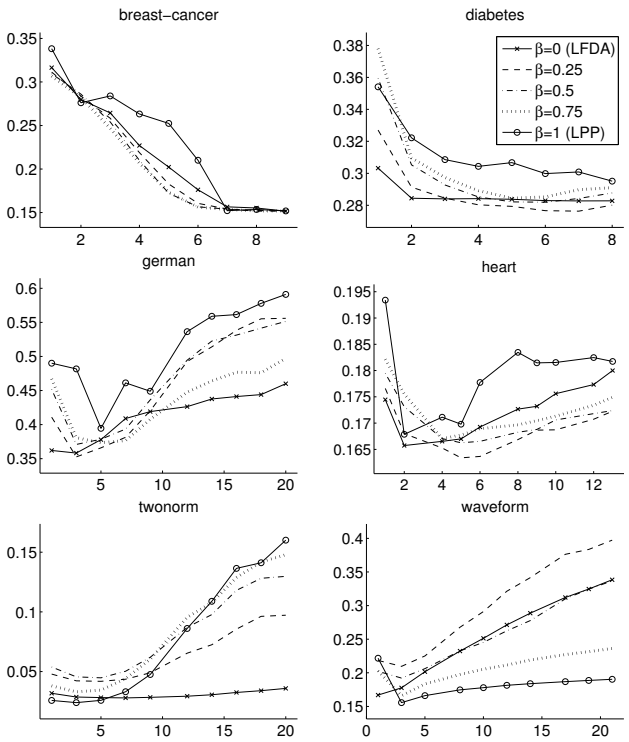

Figure 2: Misclassification error rates as functions of the dimensionality of the embedding space.

5. Conclusions

We proposed a novel semi-supervised dimensionality reduction method, which efficiently combines LFDA and LPP and includes FDA, LFDA, and LPP as special cases. The proposed method trades locality preservation with between-class separation and tends to outperform existing methods particularly when the cluster assumption is rather reliable.

The remaining issue to be discussed—which is common to all semi-supervised learning techniques—is how to optimize the values of the tuning parameters; in our case, the dimensionality r of the reduced space and the trade-off parameter β. We may employ cross-validation for this purpose, but it has two potential problems. The first problem is that number of labeled samples is usually small in semi-supervised learning scenarios and thus cross-validation is not reliable (Chapelle 06). The second problem is that the labeled samples and unlabeled samples often follow different distributions (e.g., 'two-moon' data set in Figure 1). Such a situation is called the *covariate shift* in statistics (Shimodaira 00) and standard cross-validation is known to be significantly biased. Applying covariate shift adaptation techniques to this problem would be a promising direction to be investigated.

References

[Chapelle 06] Chapelle, O., Schölkopf, B., and Zien, A. eds.: *Semi-Supervised Learning*, MIT Press, Cambridge, MA [2006]

[Fisher 36] Fisher, R. A.: The Use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics*, Vol. 7, pp. 179–188 [1936]

[Fukunaga 90] Fukunaga, K.: *Introduction to Statistical Pattern Recognition*, Academic Press, Inc., Boston, MA, 2nd edition [1990]

[He 04] He, X. and Niyogi, P.: Locality Preserving Projections, in *Advances in Neural Information Processing Systems 16*, MIT Press, Cambridge, MA [2004]

[Rätsch 01] Rätsch, G., Onoda, T., and Müller, K.-R.: Soft Margins for AdaBoost, *Machine Learning*, Vol. 42, pp. 287–320 [2001]

[Shimodaira 00] Shimodaira, H.: Improving Predictive Inference under Covariate Shift by Weighting the Log-Likelihood Function, *Journal of Statistical Planning and Inference*, Vol. 90, pp. 227–244 [2000]

[Sugiyama 06] Sugiyama, M.: Local Fisher Discriminant Analysis for Supervised Dimensionality Reduction, in *Proceedings of 23rd International Conference on Machine Learning*, pp. 905–912, Pittsburgh, PA [2006]

[Zelnik-Manor 05] Zelnik-Manor, L. and Perona, P.: Self-Tuning Spectral Clustering, in *Advances in Neural Information Processing Systems 17*, pp. 1601–1608, MIT Press, Cambridge, MA [2005]