IEICE Transactions on Information and Systems,
vol.E92-D, no.5, pp.1204–1208, 2009.

On Computational Issues of Semi-supervised Local Fisher Discriminant Analysis

Masashi Sugiyama (sugi@cs.titech.ac.jp)
Department of Computer Science, Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan

Abstract

Dimensionality reduction is one of the important preprocessing steps in practical pattern recognition. *SEmi-supervised Local Fisher discriminant analysis* (SELF)—which is a semi-supervised and local extension of Fisher discriminant analysis—was shown to work excellently in experiments. However, when data dimensionality is very high, a naive use of SELF is prohibitive due to high computational costs and large memory requirement. In this paper, we introduce computational tricks for making SELF applicable to large-scale problems.

Keywords

dimensionality reduction, semi-supervised learning, semi-supervised local Fisher discriminant analysis, sparsity, generalized eigenvalue problem

1 Introduction

Reducing dimensionality of data is one of the important challenges in pattern recognition. However, a naive use of supervised dimensionality reduction methods in high-dimensional scenarios often results in overfitting. In such cases, the use of *unlabeled samples* could be helpful [1]. A semi-supervised dimensionality reduction method called *SEmi-supervised Local Fisher discriminant analysis* (SELF) [6] has been proposed and shown to work excellently. SELF is a regularized variant of a supervised dimensionality reduction method called *Local Fisher discriminant analysis* (LFDA) [5].

An advantage of SELF in addition to its good performance is that the globally optimal solution can be obtained analytically by solving a generalized eigenvalue problem. Dimensionality of the generalized eigenvalue problem depends only on dimensionality of the feature vectors, not on the number of samples. So SELF may be applicable to a dataset with a large number of samples as long as dimensionality of the feature vectors

Table 1: Computational issues of SELF. d and n are dimensionality and the number of samples.

d	n	Sparseness	Formulation
Moderate	Large	None	Primal
Large	Moderate	None	Dual
Large	Large	Sample	Primal
Large	Large	Kernel	Dual

is moderate. However, when the feature dimensionality is very high, solving the generalized eigenvalue problem may not be computationally tractable—even more critically, the matrix which needs to be eigendecomposed cannot be stored in memory. This critical limitation makes SELF inapplicable to high-dimensional real-world problems.

The purpose of this paper is to provide computationally efficient algorithms for SELF and make it applicable to high-dimensional problems. More specifically, we introduce the following computational tricks. The first is based on the dual formulation, where the size of the generalized eigenvalue problem does not depend on the feature dimensionality but only on the number of samples. Thus this dual formulation would be computationally efficient if the number of samples is moderate. The other method makes use of the *sparsity* of feature vectors—we show that the high-dimensional *dense* matrix which needs to be eigendecomposed in SELF can be expressed by the sum of a sparse matrix and low-rank matrices. This structure allows us to efficiently solve the primal generalized eigenvalue problem even when the feature dimensionality is very high. The same trick could be applied to the dual formulation (see Table 1). Through document classification experiments, we show the effectiveness of the proposed method.

2 Dimensionality Reduction

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

2.1 Formulation

Let $\boldsymbol{x}$ ($\in \mathbb{R}^d$) be a d-dimensional sample and let $\boldsymbol{z}$ ($\in \mathbb{R}^r$) be a low-dimensional representation of $\boldsymbol{x}$, where r ($1 \leq r \leq d$) is dimensionality of the reduced space. We consider linear dimensionality reduction scenarios where an embedded representation $\boldsymbol{z}$ of a sample $\boldsymbol{x}$ is obtained by using a $d \times r$ transformation matrix $\boldsymbol{T}$ as

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

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

2.2 Principal Component Analysis (PCA)

The most well-known dimensionality reduction method would be PCA. Given samples $\{\boldsymbol{x}_i\}_{i=1}^n$, PCA seeks a transformation matrix $\boldsymbol{T}$ such that the variance of the data samples in the embedding space is maximized. Let $\boldsymbol{S}^{(\mathrm{t})}$ be the *total scatter matrix*:

$$\begin{aligned}\boldsymbol{S}^{(\mathrm{t})} :=& \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \\ =& \frac{1}{2} \sum_{i,j=1}^{n} W_{i,j}^{(\mathrm{t})} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^\top,\end{aligned}$$

where $\boldsymbol{\mu} := \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$ is the mean of all samples and $\boldsymbol{W}^{(\mathrm{t})}$ is the $n \times n$ matrix with $W_{i,j}^{(\mathrm{t})} := 1/n$. Then a PCA solution is given by the leading eigenvectors of

$$\boldsymbol{S}^{(\mathrm{t})}\boldsymbol{\varphi} = \lambda\boldsymbol{\varphi}.$$

2.3 Fisher Discriminant Analysis (FDA) for Dimensionality Reduction

A popular supervised dimensionality reduction technique is *Fisher discriminant analysis* (FDA) [3]. Suppose that we have n' labeled samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n'}$, where y_i ($\in \{1, \ldots, c\}$) is a class label associated with the sample $\boldsymbol{x}_i$ and c is the number of classes. Let n'_m be the number of labeled samples in class m.

Let $\boldsymbol{S}^{(\mathrm{b})}$ and $\boldsymbol{S}^{(\mathrm{w})}$ be the *between-class scatter matrix* and the *within-class scatter matrix*:

$$\begin{aligned}\boldsymbol{S}^{(\mathrm{b})} :=& \sum_{m=1}^{c} n'_m (\boldsymbol{\mu}_m - \boldsymbol{\mu})(\boldsymbol{\mu}_m - \boldsymbol{\mu})^\top, \\ \boldsymbol{S}^{(\mathrm{w})} :=& \sum_{m=1}^{c} \sum_{i:y_i=m} (\boldsymbol{x}_i - \boldsymbol{\mu}_m)(\boldsymbol{x}_i - \boldsymbol{\mu}_m)^\top,\end{aligned}$$

where $\boldsymbol{\mu}_m := \frac{1}{n'_m}\sum_{i:y_i=m} \boldsymbol{x}_i$ is the mean of samples in class m.

FDA seeks a transformation matrix $\boldsymbol{T}$ such that between-class scatter in the embedding space $\mathbb{R}^r$ is maximized and within-class scatter in the embedding space minimized—a solution is given by the leading generalized eigenvectors of

$$\boldsymbol{S}^{(\mathrm{b})}\boldsymbol{\varphi} = \lambda\boldsymbol{S}^{(\mathrm{w})}\boldsymbol{\varphi}.$$

$\boldsymbol{S}^{(\mathrm{b})}$ and $\boldsymbol{S}^{(\mathrm{w})}$ are related to the total scatter matrix $\boldsymbol{S}^{(\mathrm{t})}$ as

$$\boldsymbol{S}^{(\mathrm{t})} = \boldsymbol{S}^{(\mathrm{b})} + \boldsymbol{S}^{(\mathrm{w})}.$$

This can also be confirmed from the fact that $\boldsymbol{S}^{(\mathrm{b})}$ and $\boldsymbol{S}^{(\mathrm{w})}$ are expressed in the pairwise form as follows [5]:

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

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

where $\boldsymbol{W}^{(\mathrm{b})}$ and $\boldsymbol{W}^{(\mathrm{w})}$ are the $n' \times n'$ matrices with

$$W_{i,j}^{(\mathrm{b})} := \begin{cases} 1/n' - 1/n'_{y_i} & \text{if } y_i = y_j, \\ 1/n' & \text{if } y_i \neq y_j, \end{cases}$$

$$W_{i,j}^{(\mathrm{w})} := \begin{cases} 1/n'_{y_i} & \text{if } y_i = y_j, \\ 0 & \text{if } y_i \neq y_j. \end{cases}$$

Then we have

$$\boldsymbol{W}^{(\mathrm{t})} = \boldsymbol{W}^{(\mathrm{b})} + \boldsymbol{W}^{(\mathrm{w})}.$$

2.4 Local Fisher Discriminant Analysis (LFDA)

LFDA [5] is a supervised dimensionality reduction technique which is a local extension of FDA.

Let $A_{i,j}$ ($\in [0,1]$) be the *affinity* between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$; $A_{i,j}$ 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'. We assume that the affinity is symmetric, i.e., $A_{i,j} = A_{j,i}$. There are several different manners of defining the affinity; among them, we use the *local scaling heuristic* [7]:

$$A_{i,j} = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\sigma_i \sigma_j}\right).$$

$\sigma_i = \|\boldsymbol{x}_i - \boldsymbol{x}_i^{(k)}\|$ represents the local scaling around $\boldsymbol{x}_i$, where $\boldsymbol{x}_i^{(k)}$ is the k-th nearest neighbor of $\boldsymbol{x}_i$. A heuristic choice of $k = 7$ was shown to be useful through extensive simulations [7, 5].

Let $\boldsymbol{S}^{(\mathrm{lb})}$ and $\boldsymbol{S}^{(\mathrm{lw})}$ be the *local* between-class scatter matrix and the *local* within-class scatter matrix:

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

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

where $\boldsymbol{W}^{(\mathrm{lb})}$ and $\boldsymbol{W}^{(\mathrm{lw})}$ are the $n' \times n'$ matrices with[1]

$$W_{i,j}^{(\mathrm{lb})} := \begin{cases} A_{i,j}(1/n' - 1/n'_{y_i}) & \text{if } y_i = y_j, \\ 1/n' & \text{if } y_i \neq y_j, \end{cases}$$
$$W_{i,j}^{(\mathrm{lw})} := \begin{cases} A_{i,j}/n'_{y_i} & \text{if } y_i = y_j, \\ 0 & \text{if } y_i \neq y_j. \end{cases}$$

When $A_{i,j} = 1$ for all i, j (i.e., no locality), the above scatter matrices are reduced to the ones used in original FDA. Thus, LFDA could be regarded as a 'localized' variant of FDA since the effect of data pairs in the same class but having small affinity values are deemphasized in the scatter matrices.

LFDA seeks a transformation matrix $\boldsymbol{T}$ such that local between-class scatter in the embedding space is maximized and local within-class scatter in the embedding space is minimized—a solution is given by the leading generalized eigenvectors of

$$\boldsymbol{S}^{(\mathrm{lb})}\boldsymbol{\varphi} = \lambda \boldsymbol{S}^{(\mathrm{lw})}\boldsymbol{\varphi}.$$

LFDA is advantageous over original FDA in the following two respects. FDA is known to perform poorly when within-class multimodality or outliers exist [3]. On the other hand, LFDA can overcome this weakness by evaluating within-class scatter locally. The reduced dimensionality r is at most $c - 1$ in FDA [3], which is a critical limitation when the number of classes is small. On the other hand, LFDA can be generally applied to dimensionality reduction into *any* dimensional spaces, thanks to the affinity factor $A_{i,j}$.

2.5 Semi-supervised LFDA (SELF)

LFDA (and any other supervised methods) tends to suffer from overfitting when the number of samples is small. For mitigating the overfitting problem, we assume the availability of unlabeled samples; from here on, we consider the *semi-supervised* setup [1] where, in addition to the labeled samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n'}$, unlabeled samples $\{\boldsymbol{x}_i\}_{i=n'+1}^{n}$ are available.

With the help of unlabeled samples, the overfitting problem of LFDA could be mitigated by combining it with PCA—SELF smoothly bridges LFDA and PCA [6]. More specifically, a SELF solution is given by the leading generalized eigenvectors of

$$\boldsymbol{S}^{(\mathrm{rlb})}\boldsymbol{\varphi} = \lambda \boldsymbol{S}^{(\mathrm{rlw})}\boldsymbol{\varphi}, \tag{1}$$

where $\boldsymbol{S}^{(\mathrm{rlb})}$ and $\boldsymbol{S}^{(\mathrm{rlw})}$ are *regularized* local between-class scatter matrix and *regularized*

[1] A more compact expression would be

$$W_{i,j}^{(\mathrm{lb})} = \overline{A}_{i,j}/n' - W_{i,j}^{(\mathrm{lw})},$$
$$W_{i,j}^{(\mathrm{lw})} = \delta_{y_i,y_j}\overline{A}_{i,j}/n'_{y_i},$$

where $\delta_{i,j}$ denotes Kronecker's delta and $\overline{A}_{i,j} = A_{i,j}$ if $y_i = y_j$ and $\overline{A}_{i,j} = 1$ otherwise.

local within-class scatter matrix defined by

$$\boldsymbol{S}^{(\mathrm{rlb})} := (1-\beta)\boldsymbol{S}^{(\mathrm{lb})} + \beta\boldsymbol{S}^{(\mathrm{t})},$$
$$\boldsymbol{S}^{(\mathrm{rlw})} := (1-\beta)\boldsymbol{S}^{(\mathrm{lw})} + \beta\boldsymbol{I}_d.$$

$\boldsymbol{I}_d$ is the d-dimensional identity matrix. β ($\in [0,1]$) is a trade-off parameter that controls the 'degree' of unsupervisedness—when $\beta = 0$, SELF is fully supervised and is reduced to LFDA; when $\beta = 1$, SELF is fully unsupervised and is reduced to PCA; otherwise, SELF inherits the properties of both LFDA and PCA.

It is practically useful to down-weight the eigenvectors according to their associated eigenvalues [6]; more specifically, the transformation matrix is given by

$$\boldsymbol{T}^{(\mathrm{SELF})} = (\sqrt{\lambda_1}\boldsymbol{\varphi}_1|\cdots|\sqrt{\lambda_r}\boldsymbol{\varphi}_r),$$

where $\{\boldsymbol{\varphi}_k\}_{k=1}^d$ are the generalized eigenvectors associated with the generalized eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$ of Eq.(1); $\{\boldsymbol{\varphi}_k\}_{k=1}^d$ are normalized as

$$\boldsymbol{\varphi}_k^\top \boldsymbol{S}^{(\mathrm{rlw})}\boldsymbol{\varphi}_k = 1 \quad \text{for } k = 1,\ldots,d.$$

SELF was shown to be useful even when only a small number of labeled samples are available [6].

3 Computational Issues of SELF for Large-scale and High-dimensional Datasets

Here, we address computational issues of SELF.

3.1 Primal Formulation for Moderate Dimensionality

Dimensionality of the generalized eigenvalue problem (1) depends only on dimensionality of feature vectors, not on the number of samples. Thus SELF may be applicable to a dataset with a large number of samples as long as dimensionality of the feature vectors is moderate. However, when the feature dimensionality is very high, solving the generalized eigenvalue problem may not be computationally tractable—even more critically, the matrix which needs to be eigendecomposed cannot be stored in memory. This critical limitation makes SELF inapplicable to high-dimensional problems.

In the rest of this section, we show how the computational problem of SELF could be alleviated.

3.2 Dual Formulation for Moderate Sample Size

When d is very large but n is moderate, the SELF solution can be computed efficiently using the dual formulation.

For obtaining the dual problem, let us first express the primal problem (1) in terms of the sample matrix

$$\boldsymbol{X} = (\boldsymbol{x}_1|\cdots|\boldsymbol{x}_n)^\top.$$

Let $\boldsymbol{W}^{(\mathrm{rlb})}$ be the $n \times n$ matrix with

$$W_{i,j}^{(\mathrm{rlb})} := \begin{cases} (1-\beta)A_{i,j}(1/n' - 1/n'_{y_i}) + \beta/n & \text{if } y_i = y_j, \\ (1-\beta)/n' + \beta/n & \text{if } y_i \neq y_j, \\ \beta/n & \text{otherwise,} \end{cases}$$

and let $\boldsymbol{D}^{(\mathrm{rlb})}$ and $\boldsymbol{D}^{(\mathrm{lw})}$ be the $n \times n$ diagonal matrices with

$$D_{i,i}^{(\mathrm{rlb})} := \sum_{j=1}^{n} W_{i,j}^{(\mathrm{rlb})},$$
$$D_{i,i}^{(\mathrm{lw})} := \sum_{j=1}^{n} W_{i,j}^{(\mathrm{lw})}.$$

Let

$$\boldsymbol{L}^{(\mathrm{rlb})} := \boldsymbol{D}^{(\mathrm{rlb})} - \boldsymbol{W}^{(\mathrm{rlb})},$$
$$\boldsymbol{L}^{(\mathrm{rlw})} := (1-\beta)(\boldsymbol{D}^{(\mathrm{lw})} - \boldsymbol{W}^{(\mathrm{lw})}) + \beta(\boldsymbol{X}^\top\boldsymbol{X})^\dagger,$$

where † denotes the Moore-Penrose generalized inverse. Then the primal problem (1) can be expressed as follows [6][2]:

$$\boldsymbol{X}\boldsymbol{L}^{(\mathrm{rlb})}\boldsymbol{X}^\top\boldsymbol{\varphi} = \lambda\boldsymbol{X}\boldsymbol{L}^{(\mathrm{rlw})}\boldsymbol{X}^\top\boldsymbol{\varphi}. \tag{2}$$

Since $\boldsymbol{X}^\top\boldsymbol{\varphi}$ belongs to the range of $\boldsymbol{X}^\top$, it can be expressed by using some vector $\boldsymbol{\alpha}$ ($\in \mathbb{R}^n$) as

$$\boldsymbol{X}^\top\boldsymbol{\varphi} = \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\alpha} = \boldsymbol{K}\boldsymbol{\alpha},$$

where $\boldsymbol{K}$ is the $n \times n$ matrix with

$$K_{i,j} := \boldsymbol{x}_i^\top\boldsymbol{x}_j.$$

Then multiplying Eq.(2) by $\boldsymbol{X}^\top$ from the left-hand side yields the following dual eigenvalue problem:

$$\boldsymbol{K}\boldsymbol{L}^{(\mathrm{rlb})}\boldsymbol{K}\boldsymbol{\alpha} = \lambda\boldsymbol{K}\boldsymbol{L}^{(\mathrm{rlw})}\boldsymbol{K}\boldsymbol{\alpha}. \tag{3}$$

Note that $\boldsymbol{K}\boldsymbol{L}^{(\mathrm{rlw})}\boldsymbol{K}$ has a simpler expression:

$$\boldsymbol{K}\boldsymbol{L}^{(\mathrm{rlw})}\boldsymbol{K} = (1-\beta)\boldsymbol{K}(\boldsymbol{D}^{(\mathrm{lw})} - \boldsymbol{W}^{(\mathrm{lw})})\boldsymbol{K} + \beta\boldsymbol{K},$$

[2]This eigenvalue problem would be indeterminate due to rank deficiency. However, this is not a problem since we only work in the range of $\boldsymbol{X}^\top$.

so the Moore-Penrose generalized inverse of $\boldsymbol{X}^\top \boldsymbol{X}$ does not have to be computed. Let $\{\boldsymbol{\alpha}_k\}_{k=1}^d$ be the eigenvectors associated with the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$ of Eq.(3), which are normalized as

$$\boldsymbol{\alpha}_k^\top \boldsymbol{K}\boldsymbol{L}^{(\mathrm{rlw})}\boldsymbol{K}\boldsymbol{\alpha}_k = 1 \quad \text{for } k = 1, \ldots, d.$$

Then the embedded representation $\boldsymbol{z}$ of a sample $\boldsymbol{x}$ can be computed in terms of $\{\boldsymbol{\alpha}_k\}_{k=1}^r$ as

$$\boldsymbol{z} = (\sqrt{\lambda_1}\boldsymbol{\alpha}_1 | \cdots | \sqrt{\lambda_r}\boldsymbol{\alpha}_r)^\top (\boldsymbol{x}_1^\top \boldsymbol{x}, \ldots, \boldsymbol{x}_n^\top \boldsymbol{x})^\top.$$

Since Eq.(3) is an n-dimensional eigenvalue problem which is independent of d, it may be solved efficiently even when d is large as long as n is moderate.

Note that the dual formulation does not directly include the feature vectors, but only through their inner product $\boldsymbol{x}^\top \boldsymbol{x}'$. This means that SELF can be non-linearized by replacing the inner product $\boldsymbol{x}^\top \boldsymbol{x}'$ with a reproducing kernel $K(\boldsymbol{x}, \boldsymbol{x}')$ [6]. Furthermore, the use of kernels for structured data such as sequences, trees, and graphs allows us to reduce dimensionality in the structured domain.

3.3 Primal Formulation with Sparse Data

If $\boldsymbol{X}$ is sparse (i.e., only a small number of elements are non-zero), we may solve the primal eigenvalue problem efficiently even when n and d are both large.

$\boldsymbol{S}^{(\mathrm{rlb})}$ and $\boldsymbol{S}^{(\mathrm{rlw})}$ can be expressed as

$$\begin{aligned}
\boldsymbol{S}^{(\mathrm{rlb})} &= \boldsymbol{X}\boldsymbol{D}^{(\mathrm{rlb})}\boldsymbol{X}^\top - \boldsymbol{X}\boldsymbol{W}^{(\mathrm{rlb})}\boldsymbol{X}^\top, \\
\boldsymbol{S}^{(\mathrm{rlw})} &= (1-\beta)\boldsymbol{X}\boldsymbol{D}^{(\mathrm{lw})}\boldsymbol{X}^\top - (1-\beta)\boldsymbol{X}\boldsymbol{W}^{(\mathrm{lw})}\boldsymbol{X}^\top + \beta\boldsymbol{I}_d.
\end{aligned}$$

If $\boldsymbol{X}$ is sparse, then $\boldsymbol{S}^{(\mathrm{rlw})}$ is also sparse since $\boldsymbol{D}^{(\mathrm{lw})}$, $\boldsymbol{W}^{(\mathrm{lw})}$, and $\boldsymbol{I}_d$ are all sparse. On the other hand, $\boldsymbol{D}^{(\mathrm{rlb})}$ is sparse but $\boldsymbol{W}^{(\mathrm{rlb})}$ is dense, so $\boldsymbol{S}^{(\mathrm{rlb})}$ is not sparse. However, $\boldsymbol{S}^{(\mathrm{rlb})}$ has nice structure as explained below.

We can show that $\boldsymbol{W}^{(\mathrm{rlb})}$ is expressed as

$$\boldsymbol{W}^{(\mathrm{rlb})} = \boldsymbol{H} + \{(1-\beta)/n'\}\boldsymbol{\eta}\boldsymbol{\eta}^\top + (\beta/n)\mathbf{1}\mathbf{1}^\top, \tag{4}$$

where $\boldsymbol{H}$ is the $n \times n$ matrix with

$$H_{i,j} := \begin{cases} (1-\beta)\{A_{i,j}(1/n' - 1/n'_{y_i}) - 1/n'\} & \text{if } y_i = y_j, \\ 0 & \text{otherwise.} \end{cases}$$

$\boldsymbol{\eta}$ is the n-dimensional vector with $\eta_i := 1$ for $1 \leq i \leq n'$ (i.e., $\boldsymbol{x}_i$ is labeled) and $\eta_i := 0$ otherwise. $\mathbf{1}$ denotes the vector with all ones. Since the number of labeled samples is usually small in the semi-supervised setup (i.e., $n' \ll n$), $\boldsymbol{H}$ would be sparse; furthermore, if samples are sorted according to the labels, $\boldsymbol{H}$ becomes block-diagonal. Let us denote

the m-th 'diagonal-block' matrix by $\boldsymbol{H}_m$, which are dense but small. Then $\boldsymbol{S}^{(\mathrm{rlb})}$ can be expressed as

$$\begin{aligned}\boldsymbol{S}^{(\mathrm{rlb})} =& \boldsymbol{X}\boldsymbol{D}^{(\mathrm{rlb})}\boldsymbol{X}^\top - \sum_{m=1}^{c} \boldsymbol{X}_m \boldsymbol{H}_m \boldsymbol{X}_m^\top \\ & - (1-\beta)n'\boldsymbol{\xi}\boldsymbol{\xi}^\top - \beta n \boldsymbol{\mu}\boldsymbol{\mu}^\top,\end{aligned}$$

where $\boldsymbol{\xi} := \frac{1}{n'}\sum_{i=1}^{n'} \boldsymbol{x}_i$ is the mean of labeled samples; $\boldsymbol{X}_m$ is a matrix of samples in class m, which is sparse. Thus $\boldsymbol{S}^{(\mathrm{rlb})}$ is the sum of a sparse matrix and low-rank matrices. This structure is highly useful since $\boldsymbol{S}^{(\mathrm{rlb})}\boldsymbol{\varphi}$ can be computed efficiently—for example, the *eigs* function in MATLAB® that sequentially solves a sparse eigenvalue problem can take advantage of this structure for improving computational efficiency.

3.4 Dual Formulation with Sparse Kernel

The above computational trick could be applied also to the dual formulation (3). Thus, when the kernel matrix $\boldsymbol{K}$ is sparse, the dual eigenvalue problem may be solved efficiently even when both n and d are large.

4 Experiments

We have made SELF applicable to high-dimensional problems. Here, we apply it to real-world document classification tasks and evaluate its performance.

We use the *Technion Repository of Text Categorization*[3] (TechTC), which contains 295 binary document classification tasks. Each task contains a few hundred documents with category labels. We use *term frequency/inverse document frequency* (TFIDF) features for classification [4], whose dimensionality ranges from thousands to tens of thousands (their dimensionality varies depending on the task since we removed entries which are zero for all documents in the dataset). The number of document samples is relatively small in this dataset, so we use the dual formulation for computing the SELF solution. Note that the original primal formulation cannot be employed for this experiment due to its high dimensionality.

We compare the performance of 'Plain' (without dimensionality reduction), LFDA, PCA, and SELF with $\beta = 0.5$. In each method, dimensionality of the reduced space r is chosen by 5-fold cross-validation from $\{1, \ldots, 10\}$. For each dataset, we consider 3 configurations with different degrees of semi-supervisedness; given n document samples, we randomly choose $10\%, 50\%, 90\%$ of them as labeled training samples and the rest are treated as unlabeled samples. We employ the 1-nearest neighbor method for evaluating the classification accuracy of the unlabeled samples. For each dataset and each training

[3]The dataset is available from 'http://techtc.cs.technion.ac.il/techtc300/techtc300.html'. See [2] for its specification.

Table 2: Document classification results. The mean and standard deviation of the misclassification rate over 295 datasets for 100 runs are described in the row of 'Error'. For each dataset, the best method and comparable ones in terms of the mean misclassification rate over 100 runs based on the *t-test* at the significance level 5% are described in bold face. The number of times the corresponding method is judged to be the best over 295 datasets is described in the row of 'Bests'. The mean value of the reduced dimensionality r chosen by 5-fold cross validation is described in the row of 'r'.

n'/n		Plain	LFDA	SELF	PCA
0.1	Error	21.8 (2.6)	23.9 (2.6)	**19.4 (2.4)**	20.4 (2.0)
	Bests	91/295	17/295	239/295	122/295
	r	—	2.0	2.6	3.7
0.5	Error	21.0 (3.6)	14.3 (1.5)	**13.2 (1.5)**	16.5 (1.6)
	Bests	10/295	125/295	230/295	43/295
	r	—	3.3	4.1	5.5
0.9	Error	21.7 (3.4)	13.5 (2.9)	**11.8 (3.2)**	15.4 (3.5)
	Bests	15/295	149/295	253/295	98/295
	r	—	3.9	4.1	6.1

sample configuration, the experiments are repeated 100 times by changing the random choice of training samples.

The experimental results are summarized in Table 2. The table shows that all the dimensionality reduction methods tend to perform better than 'Plain' except LFDA for $n' = 0.1n$ due to strong overfitting. So dimensionality reduction overall contributes to improving the accuracy of document classification. The performance of LFDA is improved as the number of labeled samples increases, while the accuracy of 'Plain' does not improve. The performance of PCA also improves as the number of labeled samples increases; this is counter-intuitive at a glance since PCA is unsupervised. However, the labeled samples are used for choosing the reduced dimensionality r by cross-validation, so the performance tends to be improved if a large number of labeled samples are available. But the performance improvement of LFDA as the number of labeled samples is increased is more prominent than PCA, thanks to the supervised formulation.

SELF consistently outperforms LFDA and PCA for all cases, showing that SELF can effectively use information brought by unlabeled samples and improve the classification accuracy. We also tested SELF with β chosen by cross-validation; the results were almost the same as SELF with $\beta = 0.5$ (so we omit the detail), but it required more computation time.

Overall, combining LFDA and PCA by SELF is shown to be a useful dimensionality reduction method in practical document classification tasks; SELF was made applicable to such high-dimensional problems by the computational tricks introduced in this paper.

5 Conclusions

Accurately classifying high-dimensional patterns is an important challenge in pattern recognition. A semi-supervised dimensionality reduction method called SELF was demonstrated to work excellently, but its naive implementation does not allow us to apply SELF to high-dimensional problems due to high computational costs and large memory requirement. In this paper, we introduced computational tricks for making SELF applicable to high-dimensional problems. We demonstrated the usefulness of the proposed method through high-dimensional document classification simulations.

References

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

[2] D. Davidov, E. Gabrilovich, and S. Markovitch, "Parameterized generation of labeled datasets for text categorization based on a hierarchical directory," The 27th Annual International ACM SIGIR Conference, Sheffield, UK, pp.250–257, Jul. 25–29 2004.

[3] K. Fukunaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press, Inc., Boston, 1990.

[4] T. Joachims, Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms, Kluwer Academic Publishers, Boston, 2002.

[5] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis," Journal of Machine Learning Research, vol.8, pp.1027–1061, May 2007.

[6] M. Sugiyama, T. Ide, S. Nakajima, and J. Sese, "Semi-supervised local Fisher discriminant analysis for dimensionality reduction," Advances in Knowledge Discovery and Data Mining, ed. T. Washio, E. Suzuki, K.M. Ting, and A. Inokuchi, Lecture Notes in Computer Science, vol.5012, Berlin, pp.333–344, Springer, 2008.

[7] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in Advances in Neural Information Processing Systems 17, ed. L.K. Saul, Y. Weiss, and L. Bottou, pp.1601–1608, MIT Press, Cambridge, MA, 2005.