Pattern Recognition 72 (2017) 82–92

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Fast Fisher discriminant analysis with randomized algorithms

Haishan Ye[a,*], Yujun Li[a], Cheng Chen[a], Zhihua Zhang[b]

[a] Department of Computer Science and Engineering, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai 200240, China
[b] School of Mathematical Sciences, Peking University, Beijing 100871, China

ARTICLE INFO

Article history:
Received 31 October 2016
Revised 19 March 2017
Accepted 25 June 2017
Available online 27 June 2017

Keywords:
Fisher discriminant analysis
Random projection
Random feature map

ABSTRACT

Fisher discriminant analysis is a classical method for classification and dimension reduction jointly. Regularized FDA (RFDA) and kernel FAD (KFDA) are two important variants. However, RFDA will get stuck in computational burden due to either the high dimension of data or the big number of data and KFDA has similar computational burden due to kernel operations. We propose fast FDA algorithms based on random projection and random feature map to accelerate FDA and kernel FDA. We give theoretical guarantee that the fast FDA algorithms using random projection have good generalization ability in comparison with the conventional regularized FDA. We also give a theoretical guarantee that the pseudoinverse FDA based on random feature map can share similar generalization ability with the conventional kernel FDA. Experimental results further validate that our methods are powerful.

1. Introduction

Fisher discriminant analysis (FDA) is an enduring classification method in multivariate analysis and machine learning. It has been used in many applications such as face recognition [1,2], text classification [3,4], microarray data classification [5], etc. The conventional FDA problem is to find an optimal linear transformation by minimizing the total class distance and maximizing the between class distance simultaneously. It is well known that this optimization problem can be formulated as a generalized eigen-problem [6] that involves the between-class scatter matrix and total scatter matrix of the data points. However, a standard solver requires the total scatter matrix to be nonsingular, which is usually not the case in real world applications. For example, microarray datasets which have the large data dimension but small data number regime yield a singular total scatter matrix.

To address this issue, pseudo-inverse FDA and regularized FDA (RFDA) were proposed. Besides, several two-stage approximate approaches were also proposed, such as PCA + FDA [1], QR + FDA [7] and SVD+QR+FDA [8]. However, these approaches cost much time in matrix multiplication applied to high-dimensional data problems such text classification, image recognition and microarray dataset.

Kernel techniques have been introduced into FDA to circumvent the linearity assumption, because they work by nonlinearly mapping vectors in the input space to a higher-dimensional feature space and then implementing the traditional version in the feature space. There have been many different approaches to devising kernel FDA [9–13]. However, the kernel technique is hard to scale to massive data set, because $O(n^3)$ computation complexity is necessary, where n is the number of train data.

To overcome the problem of FDA and KFDA, we resort to random projection and random feature map, respectively. Random projection is a useful tool in numerical linear algebra [14,15], image analysis [16–18], and machine learning [19], etc. Random projection is also the key factor of several fast matrix decompositions, like fast singular value decomposition [20]. Random feature map plays a significant role in kernel method [21]. It has been widely used to large-scale kernel machines [22,23].

In this paper, we first apply random projection to the $n \times p$ data matrix with $n \gg p$ or $p \gg n$, where n and p are number and dimension of training data, respectively. After random projection, the data matrix will reduce to a much smaller matrix, then FDA can be trained efficiently. And our theoretical analysis shows that random projection preserves the generalization ability of the FDA on the original training data.

Further more, to circumvent the weakness of KFDA, we map data into a high finite dimension using random feature map, then apply FDA algorithms or a fast approximate FDA algorithm such as QR + FDA to the mapped data set. Our work gives theoretical analysis that FDA using random feature map approximates kernel fisher discriminant analysis well.

Finally, our empirical study invalidates the effectiveness and efficiency of our methods.

* Corresponding author.
E-mail addresses: yhs12354123@163.com (H. Ye), liyujun145@gmail.com (Y. Li), jackchen1990@gmail.com (C. Chen), zhzhang@math.pku.edu.cn (Z. Zhang).

http://dx.doi.org/10.1016/j.patcog.2017.06.029
0031-3203/

2. Background and previous work

In this paper, we are concerned with a multi-class classification problem. Given a set of n p-dimensional data points, $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \in \mathbb{R}^p$, we assume that the $\mathbf{x}_i$ are to be grouped into c disjoint classes and that each $\mathbf{x}_i$ belongs to one and only one class. We let $\mathbf{E} = [e_{ij}]$ be an $n \times c$ indicator matrix with $e_{ij} = 1$ if input $\mathbf{x}_i$ is in class j and $e_{ij} = 0$ otherwise. For these n data points, each group has n_j data points so that $\sum_{j=1}^{c} n_j = n$. Then, we define $\mathbf{\Pi} = \mathrm{diag}(n_1, \ldots, n_c)(c \times c)$.

2.1. Notation

Throughout this paper, we let $\mathbf{I}_m$ denote the $m \times m$ identity matrix, $\mathbf{1}_n$ denote the $n \times 1$ vector of ones, $\mathbf{0}$ denote the zero vector or matrix of appropriate size, and $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ is the $n \times n$ centering matrix. For convenience, we just use $\mathbf{I}$ with appropriate size sometimes.

Let $\rho = \mathrm{rank}(\mathbf{A}) \le \min\{m, n\}$ and $k \le \rho$. The singular value decomposition (SVD) of $\mathbf{A}$ can be written as

$$\mathbf{A} = \sum_{i=1}^{\rho} \sigma_i \mathbf{u}_i \mathbf{v}_i^T = \begin{bmatrix} \mathbf{U}_k & \mathbf{U}_{k\perp} \end{bmatrix} \begin{bmatrix} \mathbf{\Sigma}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{k\perp} \end{bmatrix} \begin{bmatrix} \mathbf{V}_k^T \\ \mathbf{V}_{k\perp}^T \end{bmatrix},$$

where $\mathbf{U}_k$ $(m \times k)$, $\mathbf{\Sigma}_k$ $(k \times k)$, and $\mathbf{V}_k$ $(n \times k)$ correspond to the top k singular values. We also use $\sigma_i = \sigma_i(\mathbf{A})$ to denote the ith largest singular value, $\sigma_{\max}(\mathbf{A})$ to denote the largest singular value, and $\sigma_{\min}(\mathbf{A})$ to denote the smallest nonzero singular value of $\mathbf{A}$. When $\mathbf{A}$ is symmetric positive semi-definite (SPSD), the SVD is identical to the eigenvalue decomposition, in which case we have $\mathbf{U}_\mathbf{A} = \mathbf{V}_\mathbf{A}$, $\lambda_i(\mathbf{A}) = \sigma_i(\mathbf{A})$, and $\lambda_{\min}(\mathbf{A}) = \sigma_{\min}(\mathbf{A})$ where $\lambda_i(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ are i-th largest and smallest non-zero eigenvalues of $\mathbf{A}$, respectively. Besides, $\|\mathbf{A}\| \triangleq \sigma_1$ is the spectral norm and $\|\mathbf{A}\|_F = (\sum_{i,j} a_{ij}^2)^{1/2} = (\sum_i \sigma_i^2)^{1/2}$ is the Frobenius norm. The stable rank of $\mathbf{A}$ is defined as $\mathrm{sr}(\mathbf{A}) = \|\mathbf{A}\|_F^2 / \|\mathbf{A}\|^2$.

2.2. Fisher linear discriminant analysis

Suppose the input instances are partitioned into c classes which can be expressed as $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_c]$, where $\mathbf{X}_i \in \mathbb{R}^{n_i \times p}$ contains n_i instances from the ith class and $\sum_{i=1}^{c} n_i = n$. The conventional FDA is to find the optimal linear transformation $\mathbf{A} \in \mathbb{R}^{p \times q}$ that preserves the class structure in a low dimensional space as well as in the original space. That is, $\mathbf{A}$ maps each $\mathbf{x}_i$ of $\mathbf{X}$ in the p-dimensional space to a vector $\mathbf{y}_i$ in the q-dimensional space.

The within-class, between-class, and total scatter matrices are defined as follows

$$\begin{aligned} \mathbf{S}_w &= \frac{1}{n} \sum_{i=1}^{c} \sum_{\mathbf{x} \in \mathbf{X}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T, \\ \mathbf{S}_b &= \frac{1}{n} \sum_{i=1}^{c} n_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T, \\ \mathbf{S}_t &= \mathbf{S}_b + \mathbf{S}_w, \end{aligned}$$

where $\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x}_i \in \mathbf{X}_i} \mathbf{x}_i$ is the mean of the ith class and $\mathbf{m} = \frac{1}{n} \sum_{\mathbf{x}_i \in \mathbf{X}} \mathbf{x}_i$ is the mean of the whole data set.

The conventional FDA solves the following generalized eigen-problem:

$$\mathbf{S}_b \mathbf{a}_j = \lambda_j \mathbf{S}_t \mathbf{a}_j, \quad \lambda_1 \ge \lambda_2 \ge, \ldots, \ge \lambda_{q+1} = 0 \tag{1}$$

where $q \le \min(p, c-1)$ and where we refer to $\mathbf{a}_j$ as the jth discriminant direction. Eigen-problem (1) can be expressed in matrix form as follows:

$$\mathbf{S}_b \mathbf{A} = \mathbf{S}_t \mathbf{A} \mathbf{\Lambda},$$

where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_q]$ and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$. If $\mathbf{S}_t$ is nonsingular, we obtain

$$\mathbf{S}_t^{-1} \mathbf{S}_b \mathbf{A} = \mathbf{A} \mathbf{\Lambda}.$$

However, in many application such as information retrieval, face recognition and microarray analysis, $\mathbf{S}_t$ in question can be singular since the dimension p exceeds the number of data points in general.

There are two variants of the conventional FDA in the literature to handle the ill-conditioned problem that $\mathbf{S}_t$ is singular. The first variant, the pseudo-inverse method, replaces $\mathbf{S}_t^{-1}$ by $\mathbf{S}_t^{\dagger}$ and solves the following eigen-problem:

$$\mathbf{S}_t^{\dagger} \mathbf{S}_b \mathbf{A} = \mathbf{A} \mathbf{\Lambda}.$$

Note that $\mathbf{S}_t^{\dagger}$ exists and is unique. Moreover, $\mathbf{S}_t^{\dagger}$ is identical to $\mathbf{S}_t^{-1}$ whenever $\mathbf{S}_t$ is nonsingular.

The second variant is referred as the regularized fisher discriminant analysis (RFDA). It replaces $\mathbf{S}_t$ by $\mathbf{S}_t + \delta^2 \mathbf{I}_p$ and solves the following eigen-problem:

$$(\mathbf{S}_t + \delta^2 \mathbf{I}_p)^{-1} \mathbf{S}_b \mathbf{A} = \mathbf{A} \mathbf{\Lambda}. \tag{2}$$

2.3. Kernel discriminant analysis

To apply FDA to nonlinear data, many KDA algorithms have been devised by using a so-called kernel trick. The kernel method first maps the original data into a high dimensional space $\mathcal{H}$ by a nonlinear transformation $\phi : \mathbb{R}^p \to \mathcal{H}$. Typically, ϕ is explicitly unavailable and we only know a kernel function $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ such that $k(\mathbf{x}_1, \mathbf{x}_2) = \phi(\mathbf{x}_1)^T \phi(\mathbf{x}_2)$.

In the sequel, we use the tilde notation to denote vectors and matrices in the feature space. For example, the data vectors and mean vectors in the feature space are denoted as $\tilde{x}_i$ and $\tilde{m}_j$. Accordingly, $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n](n \times g)$ and $\tilde{\mathbf{M}} = [\tilde{\mathbf{m}}_1, \ldots, \tilde{\mathbf{m}}_c](c \times g)$ are the data and mean matrices in the feature space. Here g is the dimension of the feature space. Although g is possibly infinite, we here assume that it is finite but not necessarily known. Kernel discriminant analysis (KDA) seeks to solve the following generalized eigen-problem:

$$\tilde{\mathbf{S}}_b \tilde{\mathbf{A}} = \tilde{\mathbf{S}}_t \tilde{\mathbf{A}} \mathbf{\Lambda},$$

where $\tilde{\mathbf{S}}_t$ and $\tilde{\mathbf{S}}_b$ are the pooled scatter matrix and the between-class scatter matrix in $\mathcal{H}$, respectively:

$$\begin{aligned} \tilde{\mathbf{S}}_b &= \frac{1}{n} \sum_{i=1}^{c} n_i (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})^T, \\ \tilde{\mathbf{S}}_t &= \frac{1}{n} \sum_{i=1}^{n} (\tilde{\mathbf{x}}_i - \tilde{\mathbf{m}})(\tilde{\mathbf{x}}_i - \tilde{\mathbf{m}})^T. \end{aligned}$$

Similar to FDA, KDA has pseudoinverse and regularized extension version as follows

$$\begin{aligned} \tilde{\mathbf{S}}_t^{\dagger} \tilde{\mathbf{S}}_b \tilde{\mathbf{A}} &= \tilde{\mathbf{A}} \mathbf{\Lambda}, \\ (\tilde{\mathbf{S}}_t + \delta^2 \mathbf{I}_g)^{-1} \tilde{\mathbf{S}}_b \tilde{\mathbf{A}} &= \tilde{\mathbf{A}} \mathbf{\Lambda}. \end{aligned} \tag{3}$$

2.4. Regularized Fisher discriminant analysis

In [13], The eigen-problem in (2) was reformulated as

$$\mathbf{G} \mathbf{\Pi}^{-\frac{1}{2}} \mathbf{E}^T \mathbf{H} \mathbf{X} \mathbf{A} = \mathbf{A} \mathbf{\Lambda},$$

$$\mathbf{G} = (\mathbf{X} \mathbf{H} \mathbf{X}^T + \delta^2 \mathbf{I})^{-1} \mathbf{X} \mathbf{H} \mathbf{E} \mathbf{\Pi}^{-\frac{1}{2}}, \tag{4}$$

and

$$\mathbf{G} = \mathbf{X} \mathbf{H} (\mathbf{H} \mathbf{X}^T \mathbf{X} \mathbf{H} + \delta^2 \mathbf{I})^{-1} \mathbf{E} \mathbf{\Pi}^{-\frac{1}{2}}. \tag{5}$$