ORTHOGONALIZED DISCRIMINANT ANALYSIS BASED ON GENERALIZED SINGULAR VALUE DECOMPOSITION

Wei Wu and M. Omair Ahmad, Fellow, IEEE

Department of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada
w_wu@ece.concordia.ca, omair@ece.concordia.ca

ABSTRACT

Generalized singular value decomposition (GSVD) has been used for linear discriminant analysis (LDA) to solve the small sample size problem in pattern recognition. However, this algorithm may suffer from the over-fitting problem. In this paper, we propose a novel orthogonalization technique for the LDA/GSVD algorithm to address the over-fitting problem. In this technique, an orthogonalization of the basis of the discriminant subspace derived from the LDA/GSVD algorithm is carried out through an eigen-decomposition of a small size inner product matrix. It is computationally efficient when data are high dimensional. The technique is further applied to the kernelized LDA/GSVD algorithm, mGSVD-KDA, leading to a new algorithm, referred to as GSVD-OKDA. It is shown that with linear and nonlinear kernels, this new algorithm successfully overcomes the over-fitting problem of the LDA/GSVD and mGSVD-KDA algorithms. Simulation results show that the proposed algorithms provide high recognition accuracy with low computational complexity.

Index Terms— pattern recognition, pattern classification, feature extraction, face recognition

1. INTRODUCTION

The classical linear discriminant analysis (LDA) has been widely used in many pattern recognition applications [1]-[7]. The LDA is concerned with finding a discriminant subspace in which the between-class scatter of the projected samples is maximized and simultaneously the within-class is minimized. This objective is achieved by finding the non-zero eigenvectors of $S_b S_w^{-1}$, where S_b and S_w are respectively the between-class and within-class scatter matrices [1]. However, these algorithms have conspicuous limitations in that they face the so called small sample size (SSS) problem when the number of the samples is small relative to the dimension of the samples, and they fail to capture the boundaries between the nonlinear classes. In this case, S_w is not inversible. In the past, a good number of discriminant analysis methods have been proposed to address these problems. Recently, a generalized singular value decomposition has been used in LDA (LDA/GSVD) [2] to solve the SSS problem; however, this algorithm suffers from excessive computational load with possible memory overflow when the samples have a large dimension. Subsequently, the GSVD algorithm has been modified and integrated with a kernel method leading to a kernelized discriminant algorithm [3], the mGSVD-KDA algorithm, which effectively overcomes the computational complexity problem and also has the ability to classify nonlinearly distributed patterns. However, these GSVD-based algorithms suffer from the over-fitting problem, in which the derived discriminant subspace incorporates random features that are unrelated to discrimination and the impact of these random features gets amplified in the recognition process. Ye et al. [4] have proposed a method to overcome the over-fitting problem by orthogonalizing the basis of the discrimination subspace through a QR decomposition. However, the QR decomposition has a high computational complexity. Further, the technique suffers from memory overflow problem when the patterns, such as human faces, have a very high dimension. Also, the QR decomposition cannot be applied to the kernel methods.

In this paper, a new orthogonalization technique is proposed for the LDA/GSVD algorithm to address the over-fitting problem. The proposed technique is based on the orthogonalization of the basis of the discriminant subspace through an eigen-decomposition of a small size inner product matrix. Since the LDA/GSVD algorithm may run into a memory overflow problem, the proposed scheme of orthogonalization is applied to its kernelized version, namely mGSVD-KDA algorithm. The resulting algorithm can effectively deal the over-fitting problem and at the same time it is computationally efficient for high-dimensional data. Extensive computer simulations are carried out using typical pattern recognition benchmark databases with linear or nonlinear kernels to demonstrate the computational efficiency and the recognition accuracy of the proposed scheme.

2. REVIEW OF LDA/GSVD AND ITS KERNELIZATION

2.1. LDA Based on GSVD

Let a set of n m-dimensional samples x_l ($l = 1, \cdots, n$) consist of N classes with the i^{th} class having n_i samples.

Therefore, $n = \sum_{i=1}^{N} n_i$. Let $c^{(i)}$ represent the centroid of the samples of the i^{th} class and c the global centroid of all the n samples in this set. Further, let $C = \binom{H_b^T}{H_w^T}$, where $H_b = \left[\sqrt{n_1}(c^{(1)} - c), \cdots, \sqrt{n_N}(c^{(N)} - c)\right]$, and $H_w = \left[(x_1 - c^{(1)}), \cdots, (x_{n_1} - c^{(1)}), (x_{n_1+1} - c^{(2)}), \cdots, (x_{n_1+n_2} - c^{(2)}), \cdots, (x_{n-n_N+1} - c^{(N)}), \cdots, (x_n - c^{(N)})\right]$. Then, the SVD of C can be obtained as $C = P \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix} Q^T = P_1 R Q_1^T$, where R is the singular value matrix of size $k \times k$, $k = rank(C)$, P and Q are the singular vector matrices, and P_1 and Q_1 consist of the left most k columns of P and Q, respectively, and hence correspond to the range space of C. P_1 can be further partitioned as $\binom{P_{11}}{P_{12}}$, where P_{11} and P_{12}, respectively, take the first N and the last n rows of P_1. Using the SVD of P_{11}, we have $U^T P_{11} W = \Sigma_b$, where U and W are singular vector matrices and Σ_b is the singular value matrix. Partitioning W as (W_r, W_2), where the columns of W_r correspond to the range space of P_{11}, we have an optimal transformation matrix G given by

$$G = Q_1 R^{-1} W_r. \tag{1}$$

2.2. Kernelization of LDA/GSVD

A limitation of the above LDA/GSVD algorithm is the excessive computation involved with the SVD of C when the data are high dimensional. A kernelized extension of this algorithm, which is based on the modified GSVD and referred to as mGSVD-KDA algorithm [3], can effectively overcome the computational complexity problem associated with high dimensional patterns and can capture the nonlinear pattern distribution. A kernel is a nonlinear map, $\Phi : \chi \rightarrow \mathcal{F}$, $x_l \rightarrow \phi_l$, designed to map the samples x's of the input space χ into a higher f-dimensional feature space $\mathcal{F}$, in which the classes become linearly separable and a linear discriminant analysis techniques can be applied.

As in the LDA/GSVD algorithm, define $\Phi_b = [\sqrt{n_1}(\phi^{(1)} - \phi), \cdots, \sqrt{n_N}(\phi^{(N)} - \phi)]$, $\Phi_w = \left[(\phi_1 - c^{(1)}), \cdots, (\phi_{n_1} - c^{(1)}), (\phi_{n_1+1} - c^{(2)}), \cdots, (\phi_{n_1+n_2} - c^{(2)}), \cdots, (\phi_{n-n_N+1} - c^{(N)}), \cdots, (\phi_n - c^{(N)})\right]$, and $\Gamma = \binom{\Phi_b^T}{\Phi_w^T}$, where $\phi^{(i)}$ is the centroid of the ith embedding class, and ϕ the global centroid of the mapped samples in the feature space. The SVD of Γ is then given by $\Gamma = \tilde{P} \begin{pmatrix} \tilde{R} & 0 \\ 0 & 0 \end{pmatrix} \tilde{Q}^T$, where $\tilde{P}$ and $\tilde{Q}$ are orthogonal matrices, and $\tilde{R}$ is a diagonal matrix with its elements being the non-zero singular values of Γ sorted in non-increasing order. We form a symmetric matrix $\Gamma\Gamma^T = \begin{pmatrix} \Phi_b^T \Phi_b & \Phi_b^T \Phi_w \\ \Phi_w^T \Phi_b & \Phi_w^T \Phi_w \end{pmatrix}$, and a kernel matrix is constructed as $K = (k_{lh})_{l,h=1,\cdots,n}$, whose elements are inner products in the kernel feature space determined through a kernel function such that $k_{lh} = k(x_l, x_h) = \langle \phi_l, \phi_h \rangle$. Then, the sub-matrices of $\Gamma\Gamma^T$ can be expressed in terms of K as $\Phi_b^T \Phi_b = D(B - L)^T K (B - L) D$, $\Phi_w^T \Phi_w = (I - A) K (I - A)$, $\Phi_b^T \Phi_w = D(B - L)^T K (I - A)$, where $A = diag(A_1, \cdots, A_N)$, $A_i = (1/n_i)_{n_i \times n_i}$, $B = diag(B_1, \cdots, B_N)$, $B_i = (1/n_i)_{n_i \times 1}$, $D = diag(D_1, \cdots, D_N)$, $D_i = (\sqrt{n_i})_{n_i \times n_i}$, for $i = 1, \cdots, N$, $L = (1/n)_{n \times N}$, and I is an $n \times n$ identity matrix.

The eigen-decomposition of $\Gamma\Gamma^T$ generates the eigenvector matrix $\tilde{P}$ and the non-zero eigenvalue matrix $\tilde{R}$. The leftmost z columns of $\tilde{P}$, where $z = rank(\Gamma\Gamma^T)$, form the matrix $\tilde{P}_1$, and the first N rows of $\tilde{P}_1$ form the matrix $\tilde{P}_{11}$. Through SVD, $\tilde{P}_{11}$ can be decomposed as $\tilde{P}_{11} = \tilde{U} \tilde{\Sigma}_b \tilde{W}$, where $\tilde{U}$ and $\tilde{W}$ are the singular vector matrices, and $\tilde{\Sigma}_b$ is the singular value matrix.

Suppose $\tilde{Q}$ is partitioned as $\tilde{Q} = (\tilde{Q}_1, \tilde{Q}_2)$, where $\tilde{Q}_1$ and $\tilde{Q}_2$ correspond to the range space and the null space of $\Gamma\Gamma^T$, respectively. As $\Gamma = \tilde{P}_1 \tilde{R} \tilde{Q}_1^T$, we have $\tilde{Q}_1 = \Gamma^T \tilde{P}_1 \tilde{R}^{-1}$. Let matrix $\tilde{W}_v$ consist of the left most v columns of $\tilde{W}$, where $v = rank(\Phi_b^T \Phi_b)$. We then have the optimal transformation matrix $\tilde{G}$ given by

$$\tilde{G} = Q_1 \tilde{R}^{-1} \tilde{W}_v = \Gamma^T \tilde{P}_1 \tilde{R}^{-2} \tilde{W}_v = \Gamma^T \Lambda \tag{2}$$

where $\Lambda = \tilde{P}_1 \tilde{R}^{-2} \tilde{W}_v$.

3. ORTHOGONALIZATION OF THE GSVD-BASED ALGORITHMS

The GSVD-based algorithms are susceptible to the over-fitting problem. The discriminant subspace thus derived contains random features that are unrelated to discrimination. As all the eigenvectors computed in the first stage of GSVD are maintained and the eigenvectors are divided by their associated eigenvalues, the influence of the random features on the small eigenvectors gets amplified when they are divided by their associated small eigenvalues.

Normally, there are three methods addressing the over-fitting problem. The first one is the regularization [5] in which a small positive perturbation is introduced to a matrix in order to bring small changes to large eigenvalues relative to the changes to the small eigenvalues. Thus, the effect of over-fitting is reduced when the eigenvectors are divided by the eigenvalues resulting from the perturbed matrix. The optimal regulation parameter is estimated adaptively from the training samples through cross-validation, which is very time consuming. In the second method, the smaller eigenvalues and the corresponding eigenvectors are dropped [6]. Nevertheless, there is no universal criterion to determine as to how many eigenvalues can be considered small enough to be dropped. The third approach to fixing the over-fitting problem is to orthogonalize the basis of the discriminant subspace. Ye et al. [4] orthogonalize the basis through a QR decomposition of the feature vectors of the discriminant subspace. However, QR decomposition is inefficient for high dimensional data and not conductible in the kernel feature space where the dimension is infinite.

3.1. Orthogonalization of LDA/GSVD

We now propose a novel orthogonalization method to overcome the over-fitting problem of the the LDA/GSVD algorithm. The main idea of this method is to orthogonalize the basis of the discriminant subspace by means of eigen decomposition of an inner product matrix. Through orthogonalization, the basis vectors are re-scaled so that the larger eigenvectors are assigned more discrimination capacity. Thus, the over-fitting problem is overcome. This method is efficient for high dimensional data and compatible with the kernel method.

We carry out eigen-decomposition of G^TG as

$$G^TG = W_r^T R^{-2} W_r = \vartheta \pi \vartheta^T, \tag{3}$$

where W_r consists of the left r columns of W, ϑ is an orthogonal matrix and π is a diagonal matrix. Then,

$$G_o = G\vartheta\pi^{-1/2} \tag{4}$$

is the transformation matrix with its columns mutually orthogonal. Since the size of the matrix $W_r^T R^{-2} W_r$ is small, this orthogonalization step is computationally efficient.

3.2. Orthogonalization of the Kernelized Algorithm

This orthogonalization technique also applies to the kernel-based algorithm leading to the GSVD-OKDA algorithm. Although $\tilde{G}$ is implicit, the inner product $\tilde{G}^T\tilde{G}$ can be explicitly calculated and its eigen-decomposition can be found as

$$\tilde{G}^T\tilde{G} = \tilde{W}_v^T \tilde{R}^{-2} \tilde{W}_v = \tilde{\vartheta}\tilde{\pi}\tilde{\vartheta}^T, \tag{5}$$

where $\tilde{\vartheta}$ and $\tilde{\pi}$ are the eigenvector matrix and eigenvalue matrix, respectively. Then, an orthogonalized $\tilde{G}$, $\tilde{G}_o$, is obtained such that

$$\tilde{G}_o = \tilde{G}\tilde{\vartheta}\tilde{\pi}^{-1/2}. \tag{6}$$

As in the linear algorithm, the eigen-decomposition in this step is very efficient.

Given a test image x_t with its mapping in the feature space being ϕ_t, the kernel function is applied again to obtain $q_l = k(x_l, x_t) = \langle \phi_l, \phi_t \rangle$, and subsequently form the vectors,

$$\begin{aligned} Q_b &= \left[\sqrt{n_1}(q^{(1)} - q), \cdots, \sqrt{n_N}(q^{(N)} - q)\right] \\ Q_w &= \left[(q_1 - c^{(1)}), \cdots, (q_{n_1} - c^{(1)}), (q_{n_1+1} - c^{(2)}), \cdots, \right. \\ &\quad \left. (q_{n_1+n_2} - c^{(2)}), \cdots, (q_{n-n_N+1} - c^{(N)}), \cdots, (q_n - c^{(N)})\right], \end{aligned} \tag{7}$$

where $q^{(i)} = \frac{1}{n_i} \sum_{l=(n_1+\cdots+n_i)}^{(n_1+\cdots+n_i)} q_l$ and $q = \frac{1}{n}\sum_{l=1}^{n} q_l$. Since $\Gamma\phi_t = \binom{Q_b^T}{Q_w^T}$, the projection of ϕ_t on the feature vectors can be found as $w = \tilde{G}_o^T \phi_t = \tilde{\pi}^{-1/2}\tilde{\vartheta}^T \Lambda \binom{Q_b^T}{Q_w^T}$.

4. EXPERIMENTS

In this section, experiments are conducted as an empirical evaluation of performance of the proposed orthogonalized linear and kernel algorithms. In linear discrimination, the proposed GSVD-OKDA with linear kernel is compared with mGSVD-KDA with linear kernel, its orthogonalized version based on QR decomposition (QR-mGSVD-KDA), and PCA+LDA [1]. Five SSS databases, FERET, AR, ORL, Dataset1 and Dataset2, are used in this set of experiments. The kernelized versions of these algorithms, GSVD-OKDA, mGSVD-KDA, and KPCA+LDA [7], are evaluated with two large sample size (LSS) databases, Isolet and MUSK. The images of human face databases, FERET, AR and ORL, are preprocessed to move the faces to the centers of the images and are also cropped to reduce the size. Table 1 gives the summary of the databases used in our experiments. The nearest-neighbor classifier is used throughout the experiments. The two kernel functions used in our experiments are the Gaussian radial basis function (RBF) kernel, $k(x_l, x_h) = \exp\left(-\frac{||x_l - x_h||^2}{\sigma}\right)$, where $||\cdot||$ denotes the Euclidean 2-norm and $\sigma > 0$, and the nonhomogeneous polynomial kernel, $k(x_l, x_h) = \left(\langle x_l, x_h \rangle + 1\right)^d$, where d is a positive integer. The kernel parameters are determined through cross-validation.

Table 1. Summary of Databases

	Database	Dimension (m)	No. of classes (N)	No. of training samples(n)	No. of test samples
SSS	FERET	21504	28	168	112
	AR	17640	15	75	120
	ORL	10304	40	160	120
	Dataset1	7454	7	49	161
	Dataset2	2887	4	160	160
LSS	Isolet	617	26	728	1040
	MUSK	166	2	480	400

The simulation results of the linear algorithms on the SSS databases are shown in Table 2. Since the LDA/GSVD algorithm runs into memory overflow when the three face databases are used, we instead use mGSVD-KDA algorithm with linear kernel for comparison. For the three high-dimensional face databases, memory overflow occurs when the QR-mGSVD-KDA with linear kernel being used. For the two text databases, the execution time of the QR-mGSVD-KDA with linear kernel significantly higher than that of the proposed GSVD-OKDA algorithm with linear kernel. It can be seen that, using linear kernel, the proposed GSVD-OKDA algorithm provides higher recognition accuracy compared to that of the mGSVD-KDA algorithm with a negligible overhead of the computational time.

The simulation results of the kernelized algorithms on the two LSS databases are shown in Table 3. It can be seen that from this table all the kernelized algorithms substantially outperform the LDA algorithm in terms of recognition accuracy. It is also seen that the proposed GSVD-OKDA algorithm provides a recognition accuracy that is higher than that provided

Table 2. Recognition rate (%) and execution time (seconds) of linear algorithms with small sample size databases

Database	FERET		AR		ORL		Dataset1		Dataset2	
Linear Algorithm	Recog. Rate	Exe. Time	Recog. Rate	Exe. Time	Recog. Rate	Exe. Time	Recog. Rate	Exe. Time	Recog. Rate	Exe. Time
GSVD-OKDA†	99.3	0.84	96.8	0.47	95.9	0.84	92.5	0.18	84.4	0.29
QR-mGSVD-KDA†	Memory overflow		Memory overflow		Memory overflow		92.5	7.2	84.1	12.23
mGSVD-KDA†	95.7	0.84	95.8	0.46	91.6	0.83	92.7	0.17	83.6	0.28
LDA/GSVD	Memory overflow		Memory overflow		Memory overflow		92.7	20.06	83.4	9.78
PCA+LDA	97.8	0.79	90.7	0.45	90.8	0.70	85.3	0.19	81.2	0.33

†With linear kernel.

by the mGSVD-KDA algorithm with a little increase in the computation time.

Overall, we observe that the proposed GSVD-OKDA algorithm with linear or nonlinear kernel significantly outperforms its original linear or nonlinear algorithm. The orthogonalized algorithms are also competitive to the other algorithms in terms of recognition accuracy and the computational efficiency. The experiments suggest that the over-fitting problem encountered in the GSVD-based algorithms has been overcome significantly by the proposed algorithms with a little extra computational time.

Table 3. Recognition rates (%) and execution time (seconds) with large sample size databases

Database		Isolet		MUSK	
Algorithm		Recog. Rate	Exe. Time	Recog. Rate	Exe. Time
lnr.	LDA	86.5	79.1	87.3	30.7
Poly.	GSVD-OKDA	94.7	93.4	96.9	43.2
	mGSVD-KDA	93.1	92.5	95.4	42.4
	KPCA+LDA	91.1	95.1	91.9	40.6
RBF	GSVD-OKDA	95.0	99.6	97.5	39.5
	mGSVD-KDA	94.1	98.9	96.3	38.6
	KPCA+LDA	92.3	130.2	93.0	50.5

5. CONCLUSION AND DISCUSSION

In this paper, we have proposed an orthogonalization technique to address the over-fitting problem of the LDA/GSVD algorithm. The main idea of this technique is to orthogonalize the basis of the discriminant subspace derived from the LDA/GSVD algorithm through eigen-decomposition of a small size inner product matrix. This technique is time efficient for high dimensional data. Using this technique, we have further proposed a kernelized algorithm, GSVD-OKDA, to overcome the over-fitting problem of the kernelized LDA/GSVD algorithm, mGSVD-KDA. The new algorithm with linear kernel has been demonstrated to deal effectively with the over-fitting problem and has recognition accuracy higher than that of the mGSVD-KDA algorithm with linear kernel. It has also been shown that this algorithm with nonlinear kernel provides a recognition accuracy higher than that provided by the mGSVD-KDA algorithm.

6. REFERENCES

[1] P. N. Belhumeour, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces versus Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. on Pattern Analysis and Machine Intelligence* vol. 19, no. 7, pp. 711-720, Jul. 1997.

[2] P. Howland, M. Jeon, and H. Park, "Structure perserving dimension reduction for clustered text data based on the generalized singular value decomposition," *SIAM Journal Matrix Analsis Application*, vol. 25, no. 1, pp. 165-179, 2003.

[3] W. Wu, J. He, and J. Zhang, "A kernelized discriminant analysis algorithm based on modified generalized singular value decomposition" *ICASSP 2008*, pp. 1353-1356.

[4] J. Ye and T. Xiong, "Computational and Theoretical Analysis of Null Space and Orthogonal Linear Discriminant Analysis," *Journal of Machine Learning Research*, v7, pp. 1183-1204, 2006.

[5] J. Ye, T. Xiong, Q. Li, R. Janardan, J. Bi, V. Cherkassky, and C. Kambhamettu, "Efficient model selection for regularized linear discriminant analysis," *CIKM'06*, Nov. pp. 5-11, 2006.

[6] C. Liu and H. Wechsler, "Enhanced Fisher linear discriminant models for face recognition," *Proceedings of International Conference on Pattern Recognition*, IEEE, 1998.

[7] J. Yang, A. F. Frangi, J. Yang, D. Zhang, and Z, Jin, "KPCA plus LDA: A complete kernel fisher discriminant framework for feature extraction and recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no.2, Feb. 2005.