A KERNELIZED DISCRIMINANT ANALYSIS ALGORITHM BASED ON MODIFIED GENERALIZED SINGULAR VALUE DECOMPOSITION

Wei Wu, Jijun He

Concordia University
Montreal, QC, Canada
w_wu@ece.concordia.ca, jijun_he@jmsb.concordia.ca

Jiajun Zhang

Taxes Instruments
Dallas, TX, United States
Jimmyz@ti.com

ABSTRACT

The generalized singular value decomposition based linear discriminant analysis (LDA/GSVD) algorithm has been used to solve the singularity problem faced by the traditional LDA, but it is still computationally intensive in case of high dimensional patterns; and not applicable to the nonlinearly distributed patterns. In this paper, a new kernelized discriminant analysis algorithm based on a modified GSVD is proposed. In the proposed algorithm the original input space is implicitly mapped into a higher dimensional feature space from which features are extracted by using a modified GSVD which circumvents the calculation of the large-dimension singular vectors without losing the discriminative information. The proposed algorithm solves the nonlinear distribution problem and has the advantage of being computational efficient thanks to the new feature extraction method introduced in this paper. It is shown through extensive computer simulations on the typical pattern recognition benchmark databases that the proposed algorithm outperforms the existing linear algorithms and the kernelized ones.

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

1. INTRODUCTION

The classical linear discriminant analysis (LDA) is a well-known approach for dimensionality reduction in pattern recognition. However, its application is limited due to two prevailing issues: the singularity or small sample size (SSS) issue and the nonlinear distribution issue.

A good number of LDA variants have been proposed to address the singularity issue [1]-[3]. A recent one is the LDA/GSVD algorithm [3], in which the generalized singular value decomposition (GSVD) [4] is used to solve a generalized eigenvalue problem. The application of GSVD to LDA not only provides a framework for finding the feature vectors with high recognition accuracy, but more importantly, it also relaxes the non-singularity requirement. However, this algorithm encounters excessive computational problem when the samples have a large dimension. Also, like the other linear algorithms, LDA/GSVD still cannot handle the nonlinear distribution issue.

For the nonlinear distribution problem, a remedy is to use a kernel machine [5] that linearizes the pattern distribution through a special mapping of the input samples. The integration of the kernel machine with linear discriminant methods leads to new nonlinear algorithms with enhanced recognition accuracy [6]-[9].

Applying the GSVD technique and the kernelization technique to the classical LDA at the same time can lead to a new solution that solves both of the problems [9]. In this paper, we propose a similar but different solution to what proposed by C.H. Park et al. In the proposed algorithm, referred to as mGSVD-KDA, the original input samples are nonlinearly mapped into a higher dimensional space where the pattern distribution is linearized; and a modified GSVD scheme is used to extract features in that space. The modified GSVD allows us to circumvent the calculation of the high dimensional null space of the feature vector matrix, which contains no discriminative information but requires vast computing resources in the conventional GSVD framework. The proposed algorithm overcomes the computational complexity problem associated with the high dimensional patterns and has the ability to classify nonlinearly distributed patterns, leading to an enhanced recognition accuracy. As shown by the simulation results, the proposed algorithm outperforms the KDA/GSVD algorithm in terms of recognition accuracy.

2. REVIEW OF LDA/GSVD

The objective function of the LDA

$$\xi_{opt} = \arg\max_{\xi} \frac{\xi^T S_b \xi}{\xi^T S_w \xi}, \tag{1}$$

where S_b and S_w are, respectively, the between-class and within-class scatter matrices, has no solution in the SSS situation due to the singularity of S_w. However, linear discriminant analysis based on the generalized singular value decomposition is able to find an optimal transformation matrix G that consists of ξ's even when S_w is singular.

Given a set of n m-dimensional training samples x_l, $l = 1, \cdots, n$, that consists of N classes where the i^{th} class has n_i samples, the class centroid is $c^{(i)}$, the global centroid is c, $n = \sum_{i=1}^{N} n_i$ is the sample size, and M_i is the index set $(n_1 + n_2 + \cdots + n_{i-1} + 1, \cdots\cdots, n_1 + n_2 + \cdots + n_i)$ of samples in the i^{th} class. We define

$$\begin{aligned} H_b &= \left[\sqrt{n_1}(c^{(1)} - c), \cdots, \sqrt{n_N}(c^{(N)} - c)\right], \\ H_w &= \left[(x_1 - c^{(1)}), \cdots, (x_n - c^{(N)})\right]_{l \in M_i}. \end{aligned} \tag{2}$$

Let us define a matrix $C = \binom{H_b^T}{H_w^T}$, then SVD of C can be obtained as

$$C = P \begin{pmatrix} R & 0 \\ 0 & 0 \end{pmatrix} Q^T, \tag{3}$$

where $R_{(k \times k)}$ is a diagonal matrix whose components are the non-zero singular values of C sorted in a non-increasing order, $k = rank(C)$, and $P_{((N+n)\times(N+n))}$ and $Q_{m \times m}$ are orthogonal matrices. The matrix P can be partitioned as $P = (P_1, P_2)$, where P_1 and P_2 have k and $n+N-k$ columns, respectively. 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. Now, using SVD, we can write

$$U^T P_{11} W = \Sigma_b = \begin{pmatrix} I_b & 0 & 0 \\ 0 & D_b & 0 \\ 0 & 0 & 0_b \end{pmatrix}_{(N \times k),} \tag{4}$$

$$V^T P_{12} W = \Sigma_w = \begin{pmatrix} 0_w & 0 & 0 \\ 0 & D_w & 0 \\ 0 & 0 & I_w \end{pmatrix}_{(n \times k),} \tag{5}$$

where matrix U,V and W are orthogonal matrices and Σ_b and Σ_w are diagonal matrices. In Σ_b and Σ_w, I_b and I_w are identity matrices, 0_w and 0_b are zero matrices, and D_b and D_w are diagonal matrices.

Combining (3), (4), and (5) gives

$$\begin{pmatrix} H_b^T \\ H_w^T \end{pmatrix} Q = (P_1 R, \quad 0) = \begin{pmatrix} U \Sigma_b W^T R & 0 \\ V \Sigma_w W^T R & 0 \end{pmatrix}, \tag{6}$$

Let $X = Q \begin{pmatrix} R^{-1} W & 0 \\ 0 & I \end{pmatrix}$, then (6) can be transformed into

$$H_b^T X = U(\Sigma_b \quad 0), \quad H_w^T X = V(\Sigma_w \quad 0). \tag{7}$$

From which we have,

$$X^T S_b X = \begin{pmatrix} \Sigma_b^2 & 0 \\ 0 & 0 \end{pmatrix} = D_1, X^T S_w X = \begin{pmatrix} \Sigma_w^2 & 0 \\ 0 & 0 \end{pmatrix} = D_2, \tag{8}$$

Thus, both S_b and S_w are diagonalized by matrix X. Since the null space of D_1 has little discrimination information [3], the only columns of matrix X that correspond to the range space of S_b need to be maintained during the feature extraction, and they collectively form the optimal transformation matrix G. The LDA based on this conventional GSVD may encounter excessive computation load during the SVD of C in case of high dimensional patterns.

3. KERNELIZATION OF THE MODIFIED GSVD

We now present a kernelization method that can effectively overcome the computational complexity problem associated with high dimensional patterns and capture the nonlinear pattern distribution. A kernel [5] 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 a linear discriminant analysis techniques is applied. However, the high dimensionality of the feature space makes this process computationally infeasible practically. This problem can be overcome by using the so called "kernel trick", in which the inner product of the mapped vectors in the feature space can be implicitly derived from the inner products between the input samples [5]. Since the kernel technique involving the inner product makes computation in high dimensional feature space feasible and efficient, it can also be used to overcome the computational complexity problem associated with high dimensional patterns.

Like in the LDA/GSVD algorithm, we first define

$$\begin{aligned} \Phi_b &= [\sqrt{n_1}(\phi^{(1)} - \phi), \cdots, \sqrt{n_N}(\phi^{(N)} - \phi)] \\ \Phi_w &= [(\phi_1 - \phi^{(1)}), \cdots, (\phi_n - \phi^{(N)})]_{l \in M_i}, \end{aligned} \tag{9}$$

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 given by $\Gamma = \tilde{P} \begin{pmatrix} \tilde{R} & 0 \\ 0 & 0 \end{pmatrix} \tilde{Q}^T$, where $\tilde{P}_{((N+n)\times(N+n))}$ and $\tilde{Q}_{(f \times f)}$ are orthogonal matrices, and $\tilde{R}_{(z \times z)}$ with $z = rank(\Gamma)$ is a diagonal matrix with its elements being equal to non-zero singular values of Γ sorted in a non-increasing order.

Due to the high dimensionality of Γ, it would be practically not feasible to conduct the SVD directly. Fortunately, the left singular vector matrix $\tilde{P}$ with lower dimension and singular value matrix $\tilde{R}$ can be evaluated separately by using the kernel method. We form a symmetric matrix as

$$\Gamma \Gamma^T = \begin{pmatrix} \Phi_b^T \Phi_b & \Phi_b^T \Phi_w \\ \Phi_w^T \Phi_b & \Phi_w^T \Phi_w \end{pmatrix}, \tag{10}$$

where each of the four sub-matrices is in an inner product form. The $\tilde{P}$ is exactly the eigenvector matrix of $\Gamma \Gamma^T$ and the matrix $\tilde{R}$ is the square root of its eigenvalue matrix. We can construct the kernel matrix $K = (k_{lh})_{l,h=1,\cdots,n}$ whose elements are the inner products in the feature space determined through a kernel function. Then, we can express the sub-matrices in (10) as

$$\begin{aligned} \Phi_b^T \Phi_b &= D(B - L)^T K (B - L) D \\ \Phi_w^T \Phi_w &= (I - A) K (I - A) \\ \Phi_b^T \Phi_w &= D(B - L)^T K (I - A), \end{aligned} \tag{11}$$

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 $\tilde{R}$ as

$$\Gamma\Gamma^T = \tilde{P} \begin{pmatrix} \tilde{R}^2 & 0 \\ 0 & 0 \end{pmatrix} \tilde{P}^T \tag{12}$$

The leftmost z columns of $\tilde{P}_1$, where $z = rank(\Gamma\Gamma^T)$, form the matrix $\tilde{P}_1$, and the first N rows of $\tilde{P}_1$ form the matrix $\tilde{P}_{11}$. The SVD of $\tilde{P}_{11}$ provides the orthogonal matrices $\tilde{U}$ and $\tilde{W}$ such that $\tilde{P}_{11} = \tilde{U}\tilde{\Sigma}_b\tilde{W}$.

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. Then matrix Γ can be rewritten as

$$\Gamma = (\tilde{P}_1\ \tilde{P}_2) \begin{pmatrix} \tilde{R} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \tilde{Q}_1^T \\ \tilde{Q}_2^T \end{pmatrix} = \tilde{P}_1\tilde{R}\tilde{Q}_1^T \tag{13}$$

From this equation, we have $\tilde{Q}_1 = \Gamma^T\tilde{P}_1\tilde{R}^{-1}$. Similar to defining X in Section 2, let $\tilde{X} = \tilde{Q} \begin{pmatrix} \tilde{R}^{-1}\tilde{W} & 0 \\ 0 & I \end{pmatrix}$, then, $\tilde{X}_z$, the matrix consisting of the first z columns of $\tilde{X}$ can be expressed as $\tilde{X}_z = \Gamma^T\tilde{P}_1\tilde{R}^{-2}\tilde{W}$, in which $\tilde{Q}$ is substituted and does not need to be explicitly computed. Let $\Lambda = \tilde{W}^T\tilde{R}^{-2}\tilde{P}_1^T$, we have $\tilde{X}_z^T = \Lambda\Gamma$. Further, let $\tilde{G}^T = \Lambda_v\Gamma$, where $v = rank(\Phi_b^T\Phi_b)$, and Λ_v consists of the first v rows of Λ. The columns of $\tilde{G}$ are the extracted feature vectors of the feature space.

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,

$$\begin{aligned} Q_b &= \left[\sqrt{n_1}(q^{(1)} - q), \cdots, \sqrt{n_N}(q^{(N)} - q)\right] \\ Q_w &= \left[(q_1 - q^{(1)}), \cdots, (q_n - q^{(N)})\right]_{l \in M_i}, \end{aligned} \tag{14}$$

where $q^{(i)} = \frac{1}{n_i}\sum_{l \in M_i} q_l$ 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 $w = \tilde{G}\phi_t = \Lambda_v\binom{Q_b^T}{Q_w^T}$.

4. EXPERIMENTS

Simulations are designed to compare the proposed algorithm with the existing algorithms in two categories. For linear algorithms, mGSVD-KDA with the linear kernel, LDA/GSVD, RDA [1] and PCA+LDA [2], four small sample size (SSS) databases YALE, AR, Dataset1, and Dataset2 are used. For nonlinear algorithms, mGSVD-KDA, KDA/GSVD, KPCA+LDA [7], and KRDA [8] two large sample size (LSS) databases Isolet and MUSK are used. The databases are summarized in Table 1. All the algorithms are simulated in Matlab and the nearest neighbor classifier is used for classification throughout the experiments. The simulation environment is a Pentium-4 PC with a 2.8GH CPU and 1GB RAM running on a WinXP OS.

Table 1. Summary of Databases

	Database	size	Dim.	# class	# train	# test
SSS	Yale	165	10304	15	75	90
	AR	4000	17640	15	75	120
	Dat.1	210	7454	7	49	161
	Dat.2	320	2887	4	160	160
LSS	Isolet	7797	617	26	780	1040
	MUSK	6598	166	2	500	400

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
Linear	LDA	87.5	79.1	89.3	30.7
Polyn.	mGSVD-KDA	95.2	92.5	97.9	42.4
	KDA/GSVD	94.1	131.7	97.0	49.9
	KRDA	93.8	84.0	93.8	36.6
	KPCA+LDA	93.3	95.1	92.8	40.6
RBF	mGSVD-KDA	95.5	98.9	98.0	38.6
	KDA/GSVD	94.4	144.6	97.0	52.4
	KRDA	93.4	96.1	93.5	37.9
	KPCA+LDA	92.2	130.2	93.8	50.5

The simulation results of the linear algorithms on the small sample size databases are shown in Table 2. It can be seen that as expected LDA/GSVD works well on documents classifications (Dataset1 and Dataset2), but it is very computationally expensive and it fails to handle YALE and AR due to memory overflow. The recognition accuracies of the other two algorithms are not as high as that of the proposed algorithm even though they can handle all the four cases, too.

For the kernelized algorithm comparison, two different kernel functions are employed in the simulations. One is the 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. For both of the kernel functions, the parameter d and σ were optimized for each of the algorithms during the training process such that the highest recognition accuracy for that algorithm was obtained.

The simulation results of the kernelized algorithms on the two large samples size (LSS) databases are shown in Table 3.

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

Database	YALE		AR		Dataset1		Dataset2	
Linear Algorithm	Recog. Rate	Exe. Time	Recog. Rate	Exe. Time	Recog. Rate	Exe. Time	Recog. Rate	Exe. Time
mGSVD-KDA(linear kernel mapping)	94.1	0.2	94.2	0.54	92.7	0.10	83.6	0.36
LDA/GSVD	memory overflow		memory overflow		92.7	19.45	83.4	6.31
RDA	91.6	0.19	91.1	0.79	83.5	0.12	81.3	0.34
PCA+LDA	91.6	0.19	90.5	0.58	85.3	0.16	81.2	0.54

In order to show the performance improvement of the kernelized algorithms over the traditional LDA algorithm, the simulation results of the traditional LDA algorithm are also shown in the table. It can be seen that all the kernelized algorithms substantially outperform the LDA in terms of recognition accuracy. It can also be seen that the proposed algorithm consistently outperformed all the other kernelized algorithms in terms of recognition accuracy for both of the kernel functions and for both of the databases.

5. CONCLUSION

The conventional GSVD framework has been modified and integrated with LDA leading to a new kernelized discriminant algorithm, the mGSVD-KDA algorithm. The proposed algorithm successfully overcomes the computational complexity problem of the LDA/GSVD algorithm in case of small sample size and high dimensionality and can capture the nonlinear pattern distribution. The main ideas of the new algorithm are that a nonlinear mapping is applied to transform the original input space to a higher dimensional feature space and a modified GSVD is conducted in that space, and that the calculation of the large-dimension singular vectors of SVD is circumvented without losing any discrimination information. The new algorithm with a linear kernel mapping has been demonstrated to deal effectively with the problem of high dimensionality of patterns, where the LDA/GSVD completely fails, and has competitive recognition accuracy as that of the LDA/GSVD algorithm. It has been shown that the proposed algorithm with nonlinear kernel mapping provides a recognition accuracy higher than that provided by some other existing kernelized algorithms.

6. REFERENCES

[1] J. H. Friedman, "Regularized discriminant analysis," *J. Am. Statistical Associate,* vol. 84, no. 405, pp. 957-964, 1989.

[2] P. N. Belhumeur, 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.

[3] J. Ye, R. Janardan, C. H. Park, and Haesun Park, "An optimization criterion for generalized discriminant analysis on undersampled problems," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 26, no. 8, pp. 982-994, Aug. 2004.

[4] C. C. Paige and M. A. Saunders, "Towards a Generalized Singular Value Decomposition," *SIAM Journal on Numerical Analysis,* vol. 18, no. 3, pp. 398-405, Jun. 1981.

[5] J. S. Taylor and N. Cristianini, *Kernel methods for pattern analysis.* Cambridge University Press, 2004.

[6] G. Baudat and F. Fanouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation,* vol. 12, pp. 2385-2404, 2000.

[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.

[8] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, and J, Wang, "An efficient kernel discriminant analysis method," *Pattern Recognit.*, vol. 38, no. 10, pp. 1788-1790, Oct. 2005

[9] C. H. Park and H. Park, "Nonlinear discriminant analysis using kernel functions and the generalized singular value decomposition," *SIAM Journal on Matrix Analysis and Applications,* vol. 27, no. 1, pp. 87-102, 2006.