

Generalized Nonlinear Discriminant Analysis

Li Zhang, Wei-Da Zhou, Hua Zhang, Li-Cheng Jiao

Institute of Intelligence Information Processing, Xidian University, Xi'an, China, 710071

E-mail: {zhangli, wdzhou, lchjiao}@mail.xidian.edu.cn, zhh674@163.com

Abstract

A Generalized Nonlinear Discriminant Analysis (GNDA) method is proposed, which implements Fisher discriminant analysis in a nonlinear mapping space. Linear discriminant analysis in the nonlinear mapping space corresponds to nonlinear discriminant analysis in an input space. GNDA suggests a unified framework of nonlinear discriminant analysis which includes the kernel Fisher discriminant analysis as a specific case. Experimental results on UCI data sets demonstrate the validity of our method.

1. Introduction

In pattern recognition and other data analytic tasks it is often necessary to perform feature extraction and dimensionality reduction for the high dimensionality of original data. Linear discriminant analysis (LDA), also known as Fisher linear discriminant is one of commonly used methods [1]. However LDA only extracts linear features of data, it may not work well if data have nonlinear features. To remedy this limitation of LDA, kernel Fisher discriminant (KFD), a kind of nonlinear discriminant analysis (NDA) method, has been developed for extracting nonlinear discriminant features [2]. Kernel functions used in KFD must satisfy Mercer's condition [3, 4]. There are some algorithms available for solving the optimization problem of FDA besides the classical matrix inverse plus eigendecomposition methods [11-14].

This paper proposes a generalized nonlinear discriminant analysis (GNDA) method. Similar to KFD, GNDA consists of two steps. First, data in original space are mapped into a nonlinear mapping space by using some nonlinear mapping function, and LDA is implemented in the nonlinear mapping space. Here the nonlinear mapping function can be any real-valued nonlinear function, for instance, empirical

mapping functions (or hidden functions) [5], Mercer kernel mapping [3, 4], etc. However, only the Mercer kernel mapping can be applied to KFD. As shown in Section 3 GNDA shares the features with KFD for they use the same and single Mercer kernel. In this sense the GNDA method suggests a framework which unifies the KFD method. The validity of GNDA is demonstrated by the simulation on UCI dataset.

2. LDA and FDA

For a two-class problem, let $X_1 = \{\mathbf{x}_1, \dots, \mathbf{x}_{l_1}\}$ and $X_2 = \{\mathbf{x}_1, \dots, \mathbf{x}_{l_2}\}$ be two-class sample set, respectively. Let $X = X_1 \cup X_2$. Fisher Criterion is to maximize

$$\frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}} \quad (1)$$

where \mathbf{w} is a linear transformation vector or matrix \mathbf{S}_b and \mathbf{S}_w are the between and within class scatter matrices, respectively. For two-class problems, the scatter matrices can be written as

$$\mathbf{S}_b = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \quad (2)$$

and

$$\mathbf{S}_w = \sum_{i=1}^2 \sum_{j=1, \mathbf{x}_j \in X_i}^{l_i} (\mathbf{x}_j - \mathbf{m}_i)(\mathbf{x}_j - \mathbf{m}_i)^T \quad (3)$$

where \mathbf{m}_i is the i -th class sample mean given by $(\sum_{j=1, \mathbf{x}_j \in X_i}^{l_i} \mathbf{x}_j) / l_i$. Maximizing (1) results in a generalized

eigenvalue problem $\mathbf{S}_b \mathbf{w} = \lambda \mathbf{S}_w \mathbf{w}$. KFD is a kind of NDA methods [2]. For a two-class problem, KFD has the following form

$$\max_{\alpha} \frac{\alpha^T \mathbf{S}_b' \alpha}{\alpha^T \mathbf{S}_w' \alpha} \quad (4)$$

where the quasi within-class scatter matrix $\mathbf{S}_w' = \mathbf{K} \mathbf{M} \mathbf{K}$ and the quasi between-class scatter

matrix $\mathbf{S}'_b = \sum_{m=1}^2 l_i (\mathbf{K}_m \mathbf{e}_m - \mathbf{K}\mathbf{e})(\mathbf{K}_m \mathbf{e}_m - \mathbf{K}\mathbf{e})^T$,
 $(\mathbf{K})_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) \Big|_{\mathbf{x}_i, \mathbf{x}_j \in X}$ is a kernel gram matrix,
 $(\mathbf{K}_m)_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) \Big|_{\mathbf{x}_i \in X, \mathbf{x}_j \in X_m}$, $\mathbf{e}_m = [1/l_m] \in \mathbb{R}^{l_m \times 1}$,
 $\mathbf{e} = [1/l] \in \mathbb{R}^{l \times 1}$, $\mathbf{M} = \mathbf{I} - \mathbf{N}$, $\mathbf{I} \in \mathbb{R}^{l \times l}$ is the identify
matrix, and $\mathbf{N}_{ij} = \begin{cases} 1/l_m, & \mathbf{x}_i, \mathbf{x}_j \in X_m \\ 0, & \text{otherwise} \end{cases}$. It can be
shown that $\mathbf{M}\mathbf{M} = \mathbf{M}$ and $\mathbf{M}^T = \mathbf{M}$.

3. Generalized nonlinear discriminant analysis (GNDA)

A GNDA method is presented in this section, which implements LDA in a nonlinear mapping space. The corresponding nonlinear mapping function $g(\mathbf{x})$ can be any real-valued nonlinear function. Let the set of i.i.d. patterns be

$$\{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^n, y_i \in \{1, 2, \dots, C\}, i = 1, \dots, l\}$$

where C is the total number of classes and l the number of samples. Let X_m be the m -th class sample set. Namely the whole training set is $X = \bigcup_{m=1}^C X_m$. The number of samples in X_m is denoted by l_m . thus $l = \sum_{m=1}^C l_m$. The set of the mapped patterns in the nonlinear mapping space can be expressed as

$$\{g(\mathbf{x}_i) \in \mathbb{R}^N, i = 1, \dots, l\} \quad (5)$$

where $g(\mathbf{x})$ is a pre-specified real nonlinear mapping, N is the dimensionality of the nonlinear mapping space. Let the sample matrix in the nonlinear mapping space be

$$\mathbf{G} = [g(\mathbf{x}_1), g(\mathbf{x}_2), \dots, g(\mathbf{x}_l)] \quad (6)$$

Obviously the sample matrix \mathbf{G} is a real-valued matrix with the size of $N \times l$. Since the mapped patterns in the nonlinear mapping space are definitely known once the training samples are given. The computation of any statistic about the samples in the nonlinear mapping space is feasible, such as the mean of examples which is impossible in Reproducing Kernel Hilbert Space (RKHS). In the nonlinear mapping space, scatter matrixes are defined as follows:

$$\begin{aligned} \mathbf{S}'_b &= \sum_{m=1}^C l_m (\mathbf{m}_m - \mathbf{m})(\mathbf{m}_m - \mathbf{m})^T \\ &= \sum_{m=1}^C l_i (\mathbf{G}_m \mathbf{e}_m - \mathbf{G}\mathbf{e})(\mathbf{G}_m \mathbf{e}_m - \mathbf{G}\mathbf{e})^T \end{aligned} \quad (7)$$

$$\mathbf{S}'_w = \sum_{m=1}^C \sum_{i=1, \mathbf{x}_i \in X_m}^{l_m} (g(\mathbf{x}_i) - \mathbf{m}_m)(g(\mathbf{x}_i) - \mathbf{m}_m)^T = \mathbf{G}\mathbf{M}\mathbf{G}^T \quad (8)$$

$$\mathbf{S}'_t = \mathbf{S}'_b + \mathbf{S}'_w \quad (9)$$

where \mathbf{G}_m is the m -th class sample matrix consisting of column vectors $g(\mathbf{x}_i) \Big|_{\mathbf{x}_i \in X_m}$.

Specifically, if the nonlinear mapping function takes an empirical mapping function, then the nonlinear mapping space is the empirical mapping space or the hidden space [5]. Of course, the kernel functions (is required to satisfy the symmetry only) can be used to construct the nonlinear mapping function and are not constrained to the Mercer's condition. It is known that an empirical mapping function can be constructed by a Mercer kernel

$$g(\mathbf{x}) = [k(\mathbf{x}, \mathbf{x}_1), \dots, k(\mathbf{x}, \mathbf{x}_m)]^T \quad (10)$$

or even a combination of multiple Mercer kernels

$$g(\mathbf{x}) = [k_1(\mathbf{x}, \mathbf{x}_1), \dots, k_1(\mathbf{x}, \mathbf{x}_{m_1}), k_2(\mathbf{x}, \mathbf{x}_1), \dots, k_2(\mathbf{x}, \mathbf{x}_{m_2}), \dots, k_p(\mathbf{x}, \mathbf{x}_1), \dots, k_p(\mathbf{x}, \mathbf{x}_{m_p})]^T \quad (11)$$

where $m, m_1, \dots, m_p \leq l$, k_1, k_2, \dots, k_p are any Mercer kernels.

Theorem 1: Given a nonlinear empirical mapping function $g(\mathbf{x}_i) = [k(\mathbf{x}_i, \mathbf{x}_1), \dots, k(\mathbf{x}_i, \mathbf{x}_l)]^T$, GNDA is equivalent to KFD with the same Mercer kernel $k(\mathbf{x}_i, \mathbf{x}_j)$. Namely GNDA shares the nonlinear features with KFD.

Proof: Given the nonlinear empirical mapping function $g(\mathbf{x}_i) = [k(\mathbf{x}_i, \mathbf{x}_1), \dots, k(\mathbf{x}_i, \mathbf{x}_l)]^T$, Obviously the sample matrix $\mathbf{G} = \mathbf{K}$. GNDA is to maximize

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T \left(\sum_{m=1}^C l_i (\mathbf{K}_m \mathbf{e}_m - \mathbf{K}\mathbf{e})(\mathbf{K}_m \mathbf{e}_m - \mathbf{K}\mathbf{e})^T \right) \mathbf{w}}{\mathbf{w}^T \mathbf{K} \mathbf{M} \mathbf{K} \mathbf{w}} \quad (12)$$

which is identical with the KFD problem (4). Namely $\mathbf{S}'_b = \mathbf{S}'_b$ and $\mathbf{S}'_w = \mathbf{S}'_w$. The transformation matrix \mathbf{w} of GNDA can be found by solving the following generalized eigensystem

$$\mathbf{S}'_b \mathbf{w} = \lambda \mathbf{S}'_w \mathbf{w} \quad (13)$$

Let the eigenvalues of (13) be $\{\lambda_i\}$ and the corresponding eigenvectors be $\{v_i\}$. Sort the eigenvectors v_i according to λ_i in a descend order $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{C-1} \geq \dots$. Therefore \mathbf{w} consists of the first $C-1$ eigenvectors. The nonlinear features of training samples extracted by GNDA are

$$\mathbf{Z} = \mathbf{w}^T \mathbf{K} \quad (14)$$

For KFD, its generalized eigenvalue problem is

$$\mathbf{S}'_b \boldsymbol{\alpha} = \gamma \mathbf{S}'_w \boldsymbol{\alpha} \quad (15)$$

Obviously (13) and (15) are the same problem. So if $\boldsymbol{\alpha}$ also consists of the first $C-1$ eigenvectors, then $\boldsymbol{\alpha} = \mathbf{w}$. In RKHS, there has

$$\mathbf{w}' = \Psi \boldsymbol{\alpha} \quad (16)$$

where \mathbf{w}' is the transformation matrix of KFD, Ψ is the sample matrix in RKHS, and $\mathbf{K} = \Psi^T \Psi$ holds. The nonlinear features of training samples extracted by KFD are

$$\mathbf{Z}' = \mathbf{w}'^T \Psi = \Psi^T \Psi \boldsymbol{\alpha} = \mathbf{K} \boldsymbol{\alpha} = \boldsymbol{\alpha}^T \mathbf{K} \quad (17)$$

Hence $\mathbf{Z} = \mathbf{Z}'$. This completes the proof.

Theorem 1 shows that GNDA suggests a unified framework of NDA which includes the KFD method as a specific case.

In the GNDA problem, the basic generalized eigenvalue problem is $\mathbf{S}_b \mathbf{w} = \lambda \mathbf{S}_w \mathbf{w}$. This problem can be solved by matrix inversion plus eigendecomposition, namely by applying eigendecomposition on $\mathbf{S}_w^{-1} \mathbf{S}_b$, if the scatter matrix \mathbf{S}_w is non-singular. There are many methods available to deal with the singularity problem (or under sample problem) of \mathbf{S}_w for LDA [6-9]. Through some elegant constructions those methods also can be generalized to GNDA, which will be discussed in another paper. Here we focus on coping with the singularity problem by adding a perturbation simply, similar to [2], or

$$\mathbf{S}_w = \mathbf{S}_w + \mu \mathbf{I} \quad (18)$$

where $\mu > 0$ is a very small positive constant say 10^{-8} , and \mathbf{I} is the N by N identity matrix.

4. Simulation

In order to validate the performance of GNDA, we performed some classification experiments on 10 UCI data sets [10]. Table 1 shows the attributes of these data sets.

For comparison, we also perform those with LDA and KFD methods. We set $\mu = 10^{-8}$ in all problems for three methods. After features are extracted by using these methods, classification was implemented using 1-Nearest Neighbor method which uses the standard L_2 -norm as distance measure. Especially we use empirical mapping functions as nonlinear mapping functions in GNDA. Gaussian RBF kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$ with parameter $\gamma > 0$ are used in KFD or to build the empirical mapping.

For KFD, we use Gaussian RBF kernel. For GNDA, we use two different nonlinear mapping constructions: a single Gaussian RBF kernel, and a combination of two Gaussian RBF kernels. We have to choose an

optimal Gaussian RBF kernel parameter from a discrete set $\{2^{-20}, 2^{-19}, \dots, 2^4\}$. A 10-fold crossover validation [1] is used to select the best kernel parameter and estimated the average performance. Table 2 shows the average performance of these methods.

As we can see from Table 2, the performance of KFD with Gaussian RBF kernel is identical to that of GNDA with Gaussian RBF kernel completely, which supports the Theorem 1. GNDA can adopt any real-valued nonlinear mapping function, which is its advantage over FDA. The combination of two RBF kernels can get better performance than the RBF kernel on almost all datasets except for Soy dataset. From Table 1, we know the Soy data should be a small sample dataset. Thus there has a highly singularity problem. Here we only adopt the perturbation method to avoid the singularity problem. The perturbation method used to a singularity problem can not solve the singularity problem of Soy dataset.

Table 1 10 Data Sets used in the experiments

Data set	#Attributes	#Class	Training	Testing
Breast	9	2	699	-
Cleveland heart	13	2	303	-
Ionosphere	32	2	351	-
Wpbc	33	2	198	-
Iris	4	3	150	-
Liver	6	2	345	-
Pima	8	2	768	-
Sonar	60	2	208	-
Wine	13	3	178	-
Soy	208	17	289	-

5. Conclusions

A generalized nonlinear discriminant analysis method is proposed in this paper. Any nonlinear real-valued function can be used as nonlinear mapping function in GNDA which circumvents the limitation of the Mercer permissible condition. In the experiments, the empirical nonlinear mapping which is built by the combination of two Gaussian RBF kernels is used as the nonlinear mapping. A better performance than LDA and KFD is obtained on most datasets. That is, GNDA provides another choice for the nonlinear features extraction. To obtain a better performance than KFD is possible if a suitable nonlinear mapping is used. On the other hand GNDA shares the nonlinear feature with KFD for the same kernel is used. In this sense the GNDA method unifies the KFD method.

Table 2 Performance comparison of three methods

Dataset	LDA	KFD (RBF kernel)		GNDA (RBF kernel)		GNDA (RBF + RBF kernel)	
	Error (%)	γ	Error (%)	γ	Error (%)	γ	Error (%)
Breast	4.14±3.19	2 ⁻¹⁸	3.43±2.79	2 ⁻¹⁸	3.43±2.79	(2 ⁴ , 2 ¹)	2.72±2.37
Cleveland Heart	20.83±5.22	2 ⁻¹⁶	19.96±5.29	2 ⁻¹⁶	19.96±5.29	(2 ⁴ , 2 ⁻³)	15.81±8.53
Ionosphere	18.06±4.19	2 ⁻⁹	8.24±3.72	2 ⁻⁹	8.24±3.72	(2 ¹ , 2 ⁻³)	4.17±3.00
Wpbc	21.11±8.40	2 ²	22.11±11.23	2 ²	22.11±11.23	(2 ² , 2 ⁻³)	18.17±8.51
Iris	2.67±4.66	2 ⁻⁷	2.67±3.44	2 ⁻⁷	2.67±3.44	(2 ⁻⁶ , 2 ⁻⁷)	2.00±4.50
Liver	37.86±7.42	2 ¹	31.33±4.99	2 ¹	31.33±4.99	(2 ⁻³ , 2 ⁻⁷)	29.81±6.14
Pima	31.90±4.39	2 ⁻¹⁵	28.65±5.11	2 ⁻¹⁵	28.65±5.11	(2 ⁴ , 2 ¹)	27.34±3.22
Sonar	26.29±9.40	2 ⁻¹	9.57±5.90	2 ⁻¹	9.57±5.90	(2 ⁻³ , 2 ⁴)	8.57±6.27
Wine	1.67±2.68	2 ⁻¹³	0.56±1.76	2 ⁻¹³	0.56±1.76	(2 ⁻⁷ , 2 ⁻¹⁰)	0.56±1.76
Soy	2.08±1.79	2 ⁻¹⁷	1.03±1.67	2 ⁻¹⁷	1.03±1.67	(2 ⁻⁸ , 2 ⁻¹⁰)	2.76±3.17

Here we do not discuss the singularity problem for the within-class scatter matrix, only use a simple perturbation method to avoid its singularity. From experimental results, we can see that this method works badly for highly singularity problem. The future work is to solve this problem based on [6-9]. The mathematical programming implementation for GNDA is also an aspect of the future work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 60602064, 60603019 and 60502043, and the National Research Foundation for the Doctoral Program of Higher Education of China under Grant No. 20070701022.

References

- [1] R. Duda, P. Hart, and D. Stork. *Pattern Classification*, Second edition. John Wiley & Sons, 2000.
- [2] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R., Müller. Fisher discriminant analysis with kernels. *Neural Networks for Signal Processing IX*, 1999.
- [3] C. Saunders, M. O. Stitson, J. Weston, L. Bottou, B. Schölkopf, and A. Smola. Support vector machine --- reference manual. *Technical Report CSD-TR-98-03*, Department of Computer Science, Royal Holloway, University of London, Egham, UK, 1998.
- [4] S. Saitoh. *Theory of Reproducing Kernels and Its Applications*. Longman Scientific & Technical, Harlow, England, 1988.
- [5] L. Zhang, W. Zhou, and L. Jiao. Hidden space support vector machines. *IEEE Trans. on NNs*, 15(6): 1424-1434, 2004.
- [6] P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces versus Fisher faces: Recognition using class specific linear projection. *IEEE Trans. on PAMI*, 19 (7): 711-720, 1997.
- [7] P. Howland and H. Park. Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE Trans. on PAMI*, 26(8): 995-1006, 2004.
- [8] J. Ye, and Q. Li. A two-stage linear discriminant analysis via QR-decomposition. *IEEE Trans. on PAMI*, 27(6): 929-942, 2005.
- [9] S. Zhang, and T. Sim. *Discriminant subspace analysis: A Fukunaga-Koontz approach*. *IEEE Trans. on PAMI*, 29(10): 1732-1745, 2007.
- [10] P.M. Murphy and D.W. Aha. *UCI machine learning repository*. 1992. <http://www.ics.uci.edu/~mllearn/MLRepository.html>
- [11] S. Mika, G. Rätsch, and K. Müller. A mathematical programming approach to the kernel Fisher algorithm. *NIPS2001*, 13, pp. 591-597, MIT Press, 2001.
- [12] S. Mika, A.J. Smola, and B. Schölkopf. An improved training algorithm for kernel fisher discriminants. In *Proceedings AISTATS 2001*. Morgan Kaufmann, pp. 98-104, 2001.
- [13] S.-J. Kim, A. Magnani, and S. Boyd. Optimal Kernel Selection in Kernel Fisher Discriminant Analysis. *ICML 2006*, pp. 465-472, 2006.
- [14] S.-J. Kim, A. Magnani, and S. P. Boyd. Robust Fisher Discriminant Analysis. *Advances in Neural Information Processing Systems 18*, Cambridge, MA: MIT Press address, pp. 659-666, 2006.