2010 International Conference on Pattern Recognition

An Information Theoretic Linear Discriminant Analysis Method

Haihong ZHANG, Cuntai GUAN and Kai Keng Ang

Institute for Infocomm Research, A*STAR, Singapore
Email: {hhzhang,ctguan,kkang}@i2r.a-star.edu.sg

Abstract

We propose a novel linear discriminant analysis method and demonstrate its superiority over existing linear methods. Based on information theory, we introduce a non-parametric estimate of mutual information with variable kernel bandwidth. Furthermore, we derive a gradient-based optimization algorithm for learning the optimal linear reduction vectors which maximizes the mutual information estimate. We evaluate the proposed method by running cross-validation on 2 data sets from the UCI repository, together with linear and nonlinear SVMs as classifiers. The result attests to the superority of the method over conventional LDA and its variant, aPAC.

***Keywords*-discrminant analysis, feature extraction, mutual information**

I. Introduction

Discriminant analysis (DA) consistutes a major subject in pattern recognition, and its goal is to find a low-dimensional subspace in which the class separability is maximized. The most widely-used DA method is known as Fisher linear discriminant analysis (LDA) [4], which is optimal for 2-class problems under the homoscedastic condition that all classes are Gaussian with equal covariance matrices.

The nonoptimality or sub-optimality of LDA is well-established in the literature (see [5], [10]). Specifically, neither is it able to deal with heteroscedastic data (i.e. classes do not have equal covariance matrices), nor it is Bayes optimal for $>$2-class problems even if homoscedastic condition is met. Heteroscedastic discriminant analysis (HDA) [8] was proposed that relieves LDA's homoscedasticity assumption, such that all the classes could have different covariance matrices. However, HDA was devised for uni-modal Gaussians, where data in many practical problems are more complex in nature.

In [10], the authors provided a unifying view of the criteria used by LDA, HDA, and another DA approach termed *maximum mutual information* (MMI), the last being of particular interest in this work. They introduced a hierarchy of models, from homoscedastic Gaussian model (HOG), Kumar and Andreou's heteroscedastic class-conditional Gaussian model (KAH), to a more general model called zero-information-loss model (ZIL). In this hierarchy, HOG is a special case of KAH, which is in turn a special case of ZIL. Furthermore, MMI is Bayesian optimal under ZIL.

Therefore, MMI can deal with more complex class distributions. Furthermore, MMI, which stems from information theory, naturally addresses multi-class problems. Thus, in recent years there were a few papers promoting MMI for DA [7], [6].

However, there were two limitations with the previous MMI-based DA works. First, they used Renyi's quadratic entropy in favor of its lower computational complexity, while that entropy generally deviated from Shannon's entropy. Second, the kernel bandwidth parameter was fixed, which would lead to problems in DA (see a discussion at the end of Section II).

In this paper we propose a new MMI method and demonstrate its superority over the conventional ones. The method is built up using a non-parametric (kernel-based) estimate of mutual information with Shannon's entropy. Particularly, the kernel width is variable and determined by the variable in the DA subspace. Furthermore, we derive from the estimate a gradient-based optimization algorithm.

The method is evaluated using cross-validation on 2 data sets from the UCI repository. A linear and nonlinear SVMs are used to test the separability of classes in the subspace created by the method. Conventional LDA and its variant called *approximate pairwise accuracy criterion*(aPAC) [9] were compared.

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.1016

Clearly positive results are obtained, analyzed and discussed.

II. Robust Mutual Information Estimate

Consider a sample in the original space $\mathbf{x}$. It is mapped to $\mathbf{y}$ in a subspace by a linear projection

$$\mathbf{y} = \mathbf{W}\mathbf{x}. \tag{1}$$

The projection matrix $\mathbf{W}$ comprises n_w column vectors. Denote the variable in the subspace by $\mathcal{Y}$, and the categorical variable of class labels by $\mathcal{C}$. The mutual information between $\mathcal{Y}$ and $\mathcal{C}$ is

$$\begin{aligned} I(\mathcal{Y},\mathcal{C}) &= H(\mathcal{Y}) - H(\mathcal{Y}|\mathcal{C}) \\ &= H(\mathcal{Y}) - \sum_{c\in\mathcal{C}} H(\mathcal{Y}|c)P(c), \end{aligned} \tag{2}$$

where c is a particular class label. The entropy $H(\mathcal{Y})$ is determined by p.d.f $p_y(\mathbf{y})$:

$$H(\mathcal{Y}) = -\int_{\mathbf{y}} p_y(\mathbf{y}) \ln\left(p_y(\mathbf{y})\right) d\mathbf{y}. \tag{3}$$

Like in [11], the entropy of the pattern variable $\mathcal{Y}$ can be expressed as an expectation of the function, $\ln(p(\mathbf{y}))$

$$H(\mathcal{Y}) = -E[\ln(p_{\mathbf{y}}(\mathbf{y}))] \cong -\frac{1}{n_y}\sum_{i=1}^{n_y} \ln(p_{\mathbf{y}}(\mathbf{y}_i)), \tag{4}$$

where $\mathbf{y}_i$ denotes the i-th example in the training data, $i = 1, \ldots, n_y$.

Subsequently, $p_y(\mathbf{y})$ can be estimated using kernel density estimation:

$$\hat{p}_{\mathbf{y}}(\mathbf{y}) = \frac{1}{n_y}\sum_{i=1}^{n_y} \varphi(\mathbf{y} - \mathbf{y}_i), \tag{5}$$

where

$$\varphi(\mathbf{y} - \mathbf{y}_i) = \alpha \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{y}_i)^T \mathbf{\Psi}^{-1}(\mathbf{y} - \mathbf{y}_i)\right), \tag{6}$$

Here $\mathbf{y}_i$ is a given example of the control signal variable, α is a factor that makes the integration of Eq. 5 become 1. The symbol $\mathbf{\Psi}$ denotes the bandwidth diagonal matrix of the Gaussian kernel, computed by

$$\psi_{k,k} = \zeta \frac{1}{n_y - 1}\sum_{i=1}^{n_y} (y_{ik} - \bar{y_k})^2 \tag{7}$$

where $\bar{\mathbf{y}}_k$ is the empirical mean of $\mathbf{y}_k$, the coefficient $\zeta = \left(\frac{4}{3n_z}\right)^{0.1}$ according to the normal optimal smoothing strategy [1]. Clearly, $\psi_{k,k}$ is a function of the DA linear projection W.

By substituting equation (5) into equations (4), the entropy of feature-vector variable can be estimated using

$$\hat{H}(\mathcal{Y}) = -\frac{1}{n_a}\sum_{i=1}^{n_a} \ln\left\{\frac{1}{n_a}\sum_{j=1}^{n_a} \varphi(\mathbf{y}_i - \mathbf{y}_j)\right\}, \tag{8}$$

and the conditional intra-class entropy $\hat{H}(\mathcal{Y}|c)$ can be estimated similarly by using class c examples only.

The mutual information estimate becomes

$$\hat{I}(\mathcal{Y},\mathcal{C}) = \hat{H}(\mathcal{Y}) - \sum_{c} P(c)\hat{H}(\mathcal{Y}|c) \tag{9}$$

A. Optimization Algorithm

Consider the k-th vector in the linear transformation matrix: $\mathbf{w}_k$. From Eq. 2, the gradient of mutual information estimate with respect to $\mathbf{w}_k$ is

$$\nabla_{\mathbf{w}_k} I(\mathcal{Y},\mathcal{C}) = \nabla_{\mathbf{w}_k} H(\mathcal{Y}) - \sum_{c\in\mathcal{C}} P(c)\nabla_{\mathbf{w}_k} H(\mathcal{Y}|c) \tag{10}$$

From Eq. 8, we have

$$\nabla_{\mathbf{w}_k} H(\mathcal{Y}) = -\frac{1}{n_a}\sum_{i=1}^{n_a} \beta_i \frac{1}{n_a}\sum_{j=1}^{n_a} \frac{\partial\varphi(\mathbf{y}_i - \mathbf{y}_j)}{\partial \mathbf{w}_k} \tag{11}$$

where

$$\beta_i = \left(\frac{1}{n_a}\sum_{j=1}^{n_a} \varphi(\mathbf{y}_i - \mathbf{y}_j)\right)^{-1} \tag{12}$$

From Eq. 6, we have

$$\frac{\partial\varphi(\mathbf{y}_i - \mathbf{y}_j)}{\partial \mathbf{w}_k} = -\frac{1}{2}\varphi(\mathbf{y}_i - \mathbf{y}_j)\frac{\partial(\mathbf{y}_i - \mathbf{y}_j)^T \Psi^{-1}(\mathbf{y}_i - \mathbf{y}_j)}{\partial \mathbf{w}_k} \tag{13}$$

Let's denote the quadratic function $(\mathbf{y}_i - \mathbf{y}_j)^T \Psi^{-1}(\mathbf{y}_i - \mathbf{y}_j)$ by ϑ_{ij}. And, ϑ_{ij} can be decomposed as below.

$$\vartheta_{ij} = \sum_{k_1=1}^{d_o}\sum_{k_2=1}^{d_o} \psi^{-1}_{k_1 k_2}(y_{ik_1} - y_{jk_1})(y_{ik_2} - y_{jk_2}) \tag{14}$$

The gradient of ϑ_{ij} is

$$\begin{aligned} \frac{\partial\vartheta_{ij}}{\partial \mathbf{w}_k} &= \sum_{k_1=1}^{d_o}\sum_{k_2=1}^{d_o}\left[\frac{\partial\psi^{-1}_{k_1 k_2}}{\partial \mathbf{w}_k}(y_{ik_1} - y_{jk_1})(y_{ik_2} - y_{jk_2})\right. \\ &\quad \left. + \psi^{-1}_{k_1 k_2}\frac{\partial(y_{ik_1} - y_{jk_1})(y_{ik_2} - y_{jk_2})}{\partial \mathbf{w}_k}\right] \end{aligned} \tag{15}$$

Consider that $(y_{ik_1} - y_{jk_2})^2$ is a function of $\mathbf{w}_k$ if and only if $k_1 = k$ and/or $k_2 = k$, and $\psi^{-1}_{k_1 k_2}$ is a function of $\mathbf{w}_k$ if and only if $k_1 = k_2 = k$.

Furthermore, $\psi^{-1}_{k_1k_2} = 0$ if $k_1 \neq k$ or $k_2 \neq k$. The expression of the gradient above can be written as

$$\frac{\partial \vartheta_{ij}}{\partial \mathbf{w}_k} = \frac{\partial \psi^{-1}_{kk}}{\partial \mathbf{w}_k} + \psi^{-1}_{kk_1} \frac{\partial (y_{ik} - y_{jk})^2}{\partial \mathbf{w}_k} \quad (16)$$

We further develop the above expression of gradient. To compute $\frac{\partial \psi^{-1}_{kk}}{\partial \mathbf{w}_k}$, it follows from Eq. 7 that

$$\frac{\partial \psi^{-1}_{kk}}{\partial \mathbf{w}_k} = \frac{\eta}{2} \frac{\partial \left(\mathbf{w}_k^T \Phi \mathbf{w}_k \right)}{\partial \mathbf{w}_k} = \eta \Phi \mathbf{w}_k \quad (17)$$

where

$$\eta = -2\zeta^{-1} \left(\mathbf{w}_k^T \Phi \mathbf{w}_k \right)^{-2} \quad (18)$$

and $\bar{\mathbf{x}}$ is the empirical mean of $\mathbf{x}$, and Φ the empirical covariance matrix of $\mathbf{x}$.

And

$$\frac{\partial (y_{ik} - y_{jk})^2}{\partial \mathbf{w}_k} = 2(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{w}_k \quad (19)$$

With the above equations, we can write the gradient of the mutual information estimate as below

$$\nabla_{\mathbf{w}_k} H(\mathcal{Y}) = \mathbf{A} \mathbf{w}_k \quad (20)$$

where

$$\begin{aligned} \mathbf{A} \quad = \quad & \frac{1}{2n_a^2} \sum_{i=1}^{n_a} \beta_i \sum_{j=1}^{n_a} \varphi(\mathbf{y}_i - \mathbf{y}_j) \left[\eta \Phi + \right. \\ & \left. 2\psi^{-1}_{kk} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \right] \end{aligned} \quad (21)$$

Similarly for class-conditional entropy, we have $\mathbf{A}_c$. Therefore, the gradient of the mutual information is

$$\nabla_{\mathbf{w}_k} I(\mathcal{Y}, \mathcal{C}) = \left(\mathbf{A} - \sum_c P(c) \mathbf{A}_c \right) \mathbf{w}_k \quad (22)$$

With the above equations, we are able to explicitly compute the gradient for each projection vector $\mathbf{w}_k$. We can use an iterative optimization procedure with the following updating function

$$\mathbf{w}_k^{(n_{iter}+1)} = \mathbf{w}_k^{(n_{iter})} + \lambda \nabla_{\mathbf{w}_k} I^{(n_{iter})}(\mathcal{Y}, \mathcal{C}) \quad (23)$$

where λ is the step size.

III. Results

We conduct an preliminary experimental study through 5 rounds of 5-fold cross-validtion for assessing the performance of the method.. Two data sets were selected from the UCI repository. See Table III for their specifications.

The cross-validation technique assesses how the results generated by the methods will generalize to an independent data set. Each round of 5-fold cross-validation involves partitioning a sample of data into 5 subsets, alternately performing the learning on one subset (called the training set), and validating the learned model on the others (aggregated as the test set). Five rounds of cross-validation are performed using different partitions of data in order to reduce variability. The partitions are randomly generated using the cross-validation function "crossvalind" in the MATLAB Bioinformatics toolbox.

Name	no. pattern	no. attribute	no. class
Wine	178	13	3
Yeast	1484	8	10

Table III. Datasets Description.

The peformance is assessed as the separability of the output features, in terms of classification accuracy by both linear and nonlinear classifiers, including a linear support vector machine (SVM-L) and a Gaussian-kernel support vector machine (SVM-G) (using the LIBSVM toolbox[2]).

Two conventional DA method, LDA and aPAC are implemented and compared with the proposed method. It is noteworthy that, since the optimization algorithm for this method depends on the initial condition, we use the projection matrix by LDA or aPAC as the initial guess and carry out optimization. Consequently, we denote the proposed method using the two initial conditions by LDA-MMILA and aPAC-MMILA, respectively. Besides, convergency investigation of the optimization algorithm is beyond the scope of this paper. Tentatively, the algorithm stops after 5 iterations.

The results are summarized in Table I and II. Regardless of the number of output dimension or the choice of classifier, the present method consistently yielded the lowest classification error, in both data sets except in one case (Dataset Wine, SVM-G, 8 dimensional subspace).

It is interesting to note that the comparative results were data dependent. For the dataset Wine, the methods except aPAC-MMILA produced quite comparable results, in view of the close mean error rates and the relatively large STD. For the dataset Yeast, MMILA methods produced significantly lower mean error rates. It is worthwhile to mention that, aPAC-MMILA was winning all the cases while aPAC itself lost to LDA. A plausible explanation is that, although LDA outperformed aPAC, it likely produced a strong local minimum that prevented the optimization procedure from evolving to a better solution. Thus, the optimization algorithm may be improved so as to overcome the local minimum problem, by using e.g. stochastic mechanism [3]).

DA Method	SVM-L				SVM-G			
	2	4	6	8	2	4	6	8
LDA	1.3(1.7)	1.6(1.8)	1.8(1.4)	1.8(1.8)	1.1(1.4)	1.0(1.4)	1.0(1.4)	**1.0(1.4)**
aPAC	5.4(4.2)	5.2(3.8)	5.1(3.7)	4.9(3.7)	6.9(5.0)	6.4(5.5)	7.0(5.7)	6.9(5.4)
LDA-MMILA	**1.2(1.6)**	**1.1(1.6)**	**1.8(1.4)**	**1.5(1.8)**	**1.0(1.4)**	0.9(1.3)	**0.9(1.6)**	1.3(1.6)
aPAC-MMILA	1.6(2.0)	1.5(1.8)	1.8(2.3)	1.6(1.8)	1.4(1.9)	**0.8(1.5)**	1.1(1.8)	1.2(2.2)

Table I. Dataset Wine. Classification error rates from 5×5 cross-validation are displayed in Mean(STD)% format. The lowest error rate in each column is in BOLD style. The numbers below the classifiers denote the dimensionality of the DA subspace.

DA Method	SVM-L				SVM-G			
	2	3	4	5	2	3	4	5
LDA	65.4(2.3)	59.5(2.1)	51.0(2.3)	44.5(2.8)	59.7(2.4)	58.9(3.3)	51.2(2.4)	45.0(1.6)
aPAC	61.7(2.2)	60.4(2.1)	57.8(2.0)	54.7(2.4)	64.2(3.2)	60.8(2.4)	57.5(2.2)	53.2(2.4)
LDA-MMILA	55.8(5.0)	52.8(3.9)	44.9(2.5)	44.2(2.5)	55.5(5.1)	54.6(3.9)	46.0(2.5)	44.4(2.5)
aPAC-MMILA	**47.7(2.2)**	**45.4(2.1)**	**43.2(3.1)**	**43.3(3.1)**	**46.0(2.1)**	**43.5(2.3)**	**41.5(2.6)**	**42.2(2.3)**

Table II. Dataset Yeast. See the caption of Table I for explanation.

IV. Conclusion

We have proposed a linear discriminant analysis method and demonstrated its superiority over LDA and aPAC through a preliminary study using 2 benchmark datasets. The favarable result can be attributed to the basis of the method, a non-parametric estimate of mutual information which is dependent upon the underlying data distributions. The preliminary study also indicate that, by running 5-iterations of the optimization algorithm (devised for maximizing the mutual information estimate), significantly higher class separability can be achieved in the resultant lower-dimensional subspace, compared with that by LDA and aPAC. Therefore, we suggest that futher research shall further improve and establish the efficacy of the mutual information estimate and the optimization algorithm, or to extend them for nonlinear discriminant analysis.

References

[1] A. W. Bowman and A. Azzalini. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford University Press, New York, 1997.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001. http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[3] M. Clerc and J. Kennedy. The particle swarm - explosion, stability, and convergence in a multidimensional complex space. *IEEE Transactions on Evolutionary Computing*, 6:58–73, 2002.

[4] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[5] O. Hamsici and A. Martinez. Bayes optimality in linear discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:647–657, 2008.

[6] K. Hild, D. Erdogmus, and J. Principe. An analysis of entropy estimators for blind source separation. *Signal Processing*, 86:182–194, 2006.

[7] K. Hild, D. Erdogmus, K. Torkkola, and J. Principe. Feature extraction using information-theoretic learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1385–1392, 2006.

[8] N. Kumar and A. Andreou. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communication*, 26:283–297, 1998.

[9] M. Loog, R. Duin, and R. Haeb-Umbach. Multiclass linear dimension reduction by weighted pairwise fisher criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:762–766, 2001.

[10] S. Petridis and S. Perantonis. On the relation between discriminant analysis and mutual information for supervised linear feature extraction. *Pattern Recognition*, 37:857–874, 2004.

[11] P. Viola and W. W. III. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24:137–154, 1997.