PERGAMON

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Pattern Recognition 37 (2004) 1545–1547

PATTERN
RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Rapid and Brief Communication

Alternative linear discriminant classifier

Songcan Chen*, Xubing Yang

Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, PR China

Received 24 October 2003; accepted 3 November 2003

Abstract

Fisher linear discriminant analysis (FLDA) finds a set of optimal discriminating vectors by maximizing Fisher criterion, i.e., the ratio of the between scatter to the within scatter. One of its major disadvantages is that the number of its discriminating vectors capable to be found is bounded from above by C-1 for C-class problem. In this paper for binary-class problem, we propose alternative FLDA to breakthrough this limitation by only replacing the original between scatter with a new scatter measure. The experimental results show that our approach give impressive recognition performances compared to both the Fisher approach and linear SVM.

Keywords: Fisher linear discriminant analysis; Feature extraction; Alternative linear discriminant classifier; Support vector machines

1. Fisher linear discriminant analysis (FLDA) and some problems

In this paper, we focus two class discriminating problem. It is well-known that FLDA is a popular feature extraction and discriminating approach [1] in pattern recognition and data analysis communities. Formally, it can briefly be formulated as follows: Given two pattern classes $X^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_{N_i}^{(i)}]$, $i = 1, 2$ with N_i D-dimensional patterns in the ith class, respectively. FLDA attempts to seek an optimal discriminating vector φ by maximizing the Fisher criterion:

$$J(\varphi) = \frac{\varphi^{\mathrm{T}} S_b \varphi}{\varphi^{\mathrm{T}} S_w \varphi}, \tag{1}$$

where S_b is the between-class scatter matrix and denoted by $S_b = (m_1 - m_2)(m_1 - m_2)^{\mathrm{T}}$ and S_w is the within-class scatter matrix and denoted by $S_w = \sum_{i=1}^{2} \sum_{j=1}^{N_i} (x_j^{(i)} - m_i)(x_j^{(i)} - m_i)^{\mathrm{T}}$, m_1 and m_2 denote two corresponding class means, respectively. By maximizing criterion (1), we can get only one optimal discriminating vector φ denoted by $S_w^{-1}(m_1 - m_2)$ due to that the rank of S_b is at most 1 for two class problem (exact one here, we call it the rank problem hereafter). It is this point that limits us to search more discriminating directions to further boost recognition performance of FLDA, which also inspires us to design a new between-class scatter measure to replace the original S_b to breakthrough the notorious limitation so that we not only can promote the recognition performance of the original FLDA classifier but also still keep its analytical simplicity.

The rest of the paper is organized as follows: Section 2 gives a formulation of our AFLDA. Section 3 briefly introduces the linear SVM for self-consistence and finally the experimental results on seven real-world data sets are exhibited in Section 4.

2. Alternative FLDA (AFLDA)

2.1. Alternative discriminant criterion and the derivation of optimal discriminating direction

In order to construct a new between-class scatter measure to overcome the rank problem, let us first review the geometrical meaning of the FLDA between-class scatter

* Corresponding author. Tel.: +86-25-489-1216; fax: +86-25-489-3777.

E-mail address: s.chen@nuaa.edu.cn (S. Chen).

0031-3203/$30.00 © 2003 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2003.11.008

measure in the projected space defined by $\phi^{\mathrm{T}}S_b\phi=|\phi^{\mathrm{T}}S_b\phi|=(\phi^{\mathrm{T}}m_1-\phi^{\mathrm{T}}m_2)^2$, it defines a squared difference between the average projected lengths of the two classes and then gets maximized, equivalently making the two classes separated as far from each other as possible. Thus, only one projection can be found. Now we construct the following new scatter measure denoted by

$$|\varphi^{\mathrm{T}}S_{nb}\varphi| = \left|\frac{1}{N_1}\sum_{i=1}^{N_1}\|\varphi^{\mathrm{T}}x_i^{(1)}\|^2 - \frac{1}{N_2}\sum_{j=1}^{N_2}\|\varphi^{\mathrm{T}}x_j^{(2)}\|^2\right|, \tag{2}$$

where $S_{nb}=(1/N_1)\sum_{i=1}^{N_1}x_i^{(1)}x_i^{(1)\mathrm{T}}-1/N_2\sum_{j=1}^{N_2}x_j^{(2)}x_j^{(2)\mathrm{T}}$, $\|z\|$ stands for the length of the vector z.

Eq. (2) defines an absolute value of the difference between the two-class average squared projected lengths. Maximizing Eq. (2) equivalently makes the projected vectors of one class have an average length greater than those of the other class, hence resulting in a good separation between the two classes. Here S_{nb} is an indefinite symmetric matrix but the upper-bound of its rank is greater than that of S_b (Claim 1) due to such a fact from the matrix theory [5] that for any matrices A, B and invertible E, rank(AEB) is equal or less than the minimum of both rank(A) and rank(B), where rank(X) is the rank of matrix X. Now we rewrite S_{nb}, in a matrix form, as XEX^{T}, where $X=[X^{(1)},X^{(2)}]$ is a given training pattern matrix and $E=\mathrm{diag}[1/N_1I_1,-1/N_2I_2]$ is invertible with I_i being N_i*N_i identity matrix ($i=1,2$). From the above fact, we have $\mathrm{rank}(XEX^{\mathrm{T}})\leqslant\min(\mathrm{rank}(X),\mathrm{rank}(X^{\mathrm{T}}))=\mathrm{rank}(X)$, therefore the above Claim 1 holds and is also confirmed by our experiments. As a consequence, we can obtain multiple discriminating projections for the two-class problem using our new between-class scatter measure represented by

$$J_n(\varphi)=\frac{|\varphi^{\mathrm{T}}S_{nb}\varphi|}{\varphi^{\mathrm{T}}S_w\varphi}, \tag{3}$$

where S_w is defined as before. By differentiating Eq. (3) and zeroing its derivative, we have the following eigen-system satisfied by a set of eigenvectors $\Phi_d=[\phi_1,\phi_2,\ldots,\phi_d]$:

$$S_w^{-1}S_{nb}\Phi_d=\Phi_d\Lambda_d, \tag{4}$$

where $\Lambda_d=\mathrm{diag}[\lambda_1,\lambda_2,\ldots,\lambda_d]$ is a set of eigenvalues of Eq. (4), but they are not non-negative anymore because of indefinite of S_{nb}. In order to ensure criterion (3) maximization, we select the first d eigenvectors with respect to the first d largest absolute eigenvalues to comprise the so-needed projection matrix Φ_d because now

$$J_n(\Phi_d)=\sum_{i=1}^{d}\frac{|\varphi_i^{\mathrm{T}}S_{nb}\varphi_i|}{\varphi_i^{\mathrm{T}}S_w\varphi_i}=\sum_{i=1}^{d}|\lambda_i|$$

determines its optimization. Where d is taken as a minimum of satisfying the inequality $\sum_{i=1}^{d}|\lambda_i|/\sum_{i=1}^{D}|\lambda_i|\geqslant\theta\ (>0)$.

2.2. Classification rules for FLDA and AFLDA

To classify an unknown input pattern x_u with FLDA and AFLDA, respectively, we first project it along the discriminating vectors calculated from training patterns and then use the following corresponding decision rules, respectively, to classify it.

For FLDA, if $(m_2-m_1)^{\mathrm{T}}\varphi\,\varphi^{\mathrm{T}}(x_u-m_0)<0$ and for AFLDA, if $(m_2-m_1)^{\mathrm{T}}\Phi_d\Phi_d^{\mathrm{T}}(x_u-m_0)<0$, then x_u is, respectively, classified as class 1, otherwise class 2, where m_0 is defined as $N_1m_1+N_2m_2/N_1+N_2$ for FLDA, and $N_1m_1+N_2m_2/N_1+N_2$ for AFLDA, respectively.

In addition, in order to compare with recently developed support vector machine (SVM), we also give a brief introduction in the following section.

3. SVMs

SVMs are based on the structural risk minimization (SRM) principle and aim at maximizing the margin between the points of the two classes by solving a convex quadratic programming problem. The solution to that problem gives us a hyper-plane having the maximum margin that is attainable between the two classes. SRM is a trade-off between the quality of the approximation of the given data and the complexity of the approximating function. Vapnik [2] showed that generalization error is bounded by a number proportional to the ratio R^2/γ^2, where R is the radius of the sphere that contains all training patterns and γ is the margin. There, in order to have a tighter bound and a better generalization, we need to reduce the radius while maximizing the margin, which is the goal of SRM principle and the SVMs.

The separating hyper-plane or decision function following the SRM can be determined by

$$f(x)=w^{\mathrm{T}}x=\sum_{i=1}^{N_1}\alpha_i^{(1)}(x^{\mathrm{T}}x_i^{(1)})-\sum_{j=1}^{N_2}\alpha_j^{(2)}(x^{\mathrm{T}}x_j^{(2)})+b, \tag{5}$$

where b is a threshold or bias term and given in Ref. [3] and $\alpha_j^{(i)}\geqslant 0\ \forall i,j$ with $\sum_{i=1}^{N_1}\alpha_i^{(1)}+\sum_{ji=1}^{N_2}\alpha_j^{(2)}=1$. And any vector $x_j^{(i)}$ with respect to $\alpha_j^{(i)}\neq 0$ is a support vector of the optimal hyper-plane. The sign of $f(x)$ determines the class membership of x. In this paper, for performing fairly comparison, we just adopt the linear SVM expressed by Eq. (5).

4. Experimental results

After formulating our new FLDA and introducing both FLDA and SVMs, we are in a position to carry out several experiments on seven real-world data sets in part from the UCI Repository [4]. Each data set was divided into two parts with varying sizes—training and testing sets. For classification problem under each given division, 20 independent runs were performed and their results (testing accuracies) are averaged. In all experiments, we set θ to 0.98 to determine the so-called number of the optimal discriminating

Table 1
Testing or generalization accuracies (%) on the testing sets

	Data sets	SVM	FLDA	AFLDA
IRIS	(60[a]/ 70[a]/ 80[a])	94.0[b]/ 94.2[b]/ 93.4[b]	86.1/ 84.7/ 87.6	95.5/ 95.1/ 97.4 (2)[c]
WINE	(80/ 90/ 100)	94.8/ 95.5/ 96.0	74.2/ 77.1/ 77.1	96.9/ 97.3/ 96.8 (5)
WBC	(250/300/350)	96.3/ 96.6/ 96.6	85.2/ 85.4/ 86.7	97.2/ 97.0/ 97.4 (8)
LIV	(80/100/120)	63.4/ 62.8/ 64.5	60.0/ 55.6 / 57.8	62.3/ 62.6 / 62.9(4)
DIAB	(80/100/120)	72.9/ 74.9/ 71.8	66.0/ 66.3/ 66.4	73.6/ 73.8/ 73.9(6)
MUSK	(200/300/500)	32.7/ 45.4/ 24.9	64.9/ 69.8/ 69.3	66.7/ 78.7/ 84.6(135)
WDBC	(50/100/200)	91.1/ 93.2/ 94.4	84.6/ 84.4/84.0	87.0 /93.3/96.0/ (17)

[a] The numbers of training samples.
[b] The testing accuracies for given partitions.
[c] The number of the optimal discriminating features found.

projections according to criterion (3). The data sets used are briefly described as follows:

(1) *Iris* (*IRIS*): Class2 (50 data) vs. Class3 (50 data) (linear non-separable), four dimension (4D) each input pattern and the size of the data set is 100;
(2) *wine in VISTA* (*WINE*): Class2 (71) vs. Class1 and Class3 (107), 13D, 178;
(3) *Wisconsin breast cancer* (*WBC*): Class1 (444) vs. Class2 (239), 9D, 683;
(4) *Bupa liver disorder* (*LIV*): Class1 (145) vs. Class2 (200), 6D, 345;
(5) *Pima Indian diabetes* (*DIAB*): Class1 (500) vs. Class2 (268), 8D, 768;
(6) *Musk clean*2(*Musk*):Class1 (5581) vs. Class2 (1017), 166D, 6598;
(7) *Wisconsin diagnosis breast* (*WDBC*): Class1 (212) vs. Class2 α(357), 30D, 569.

The calculated testing accuracies are given in Table 1 and thus equivalently, also explain the respective generalization abilities of the individual classifying algorithms. From the table, we draw a conclusion that AFLDA does breakthroughs the rank limitation of FLDA in all of our experiments here as remarked in the parenthesis of the rightmost column of Table 1 and has better generalization ability than both the linear SVM and FLDA in most of the data sets (as underlined in the Table 1). Finally, from the computational efficiency, the SVM has much heavier cost than both FLDA and AFLDA, which contributes mainly to its quadratic programming optimization. And in contrast, both the FLDA and AFLDA have almost the same computational complexities due to their direct and simple analytical solutions.

Acknowledgements

We thank National Science Foundations of China and of Jiangsu under Grant Nos. 60271017 and BK2002092, the “QingLan” Project Foundation of Jiangsu Province and the Returnee Foundation of China Scholarship Council for partial supports, respectively.

References

[1] G.J. Mclachlan, Discriminant Analysis and Statistical Pattern Recognition, Wiley, New York, 1992.

[2] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.

[3] N. Cristianini, J. Shewe-Taylor, Support Vector Machines, Cambridge University Press, Cambridge, 2000.

[4] UCI Repository available at http://www.kernel-machines.org/datasets/.

[5] R.A. Horn, C.R. Johnson, Matrix analysis, Cambridge University Press, Cambridge, 1985.