Hyperspectral Data Classification
Using Nonparametric Weighted Feature Extraction

Bor-Chen Kuo, *Member, IEEE*
Department of Mathematic Education
National Taichung Teachers College
Taichung, Taiwan 403
Email: kbc@mail.ntctc.edu.tw

David A. Landgrebe, *Life Fellow, IEEE*
School of Electrical and Computer Engineering
Purdue University, West Lafayette, Indiana 47907-1285
Email: landgreb@ecn.purdue.edu

Hyperspectral Data Classification Using Nonparametric Weighted Feature Extraction

Bor-Chen Kuo, *Member, IEEE*
Department of Mathematic Education
National Taichung Teachers College
Taichung, Taiwan 403
Email: kbc@mail.ntctc.edu.tw

David A. Landgrebe, *Life Fellow, IEEE*
School of Electrical and Computer Engineering
Purdue University, West Lafayette, Indiana 47907-1285
Email: landgreb@ecn.purdue.edu

***Abstract*-In this paper, a new nonparametric feature extraction method is proposed for high dimensional multiclass pattern recognition problems. It is based on a nonparametric extension of scatter matrices. There are at least two advantages to using the proposed nonparametric scatter matrices. First, they are generally of full rank. This provides the ability to specify the number of extracted features desired and to reduce the effect of the singularity problem. This is in contrast to parametric discriminant analysis, which usually only can extract L–1 (number of classes minus one) features. In a real situation, this may not be enough. Second, the nonparametric nature of scatter matrices reduces the effects of outliers and works well even for non-normal data sets. The new method provides greater weight to samples near the expected decision boundary. This tends to provide for increased classification accuracy.**

I. Introduction

Discriminant Analysis Feature Extraction (DAFE) is often used for dimension reduction in classification problems. It is also called the parametric feature extraction method in [1], since DAFE uses the mean vector and covariance matrix of each class. The purpose of DAFE is to find a transformation matrix A such that the class separability of transformed data (Y) is maximized. Usually within-class, between-class, and mixture scatter matrices are used to formulate the criteria of class separability. A within-class scatter matrix for L classes is expressed by [1]:

$$S_w = \sum_{i=1}^{L} P_i E\{(X - m_i)(X - m_i)^T \mid \omega_i\} = \sum_{i=1}^{L} P_i \Sigma_i = \sum_{i=1}^{L} P_i S_{wi}$$

where P_i means the prior probability of class i, m_i is the class mean and Σ_i is the class covariance matrix. A between-class scatter matrix is expressed as

$$S_b = \sum_{i=1}^{L} P_i (m_i - m_0)(m_i - m_0)^T = \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} P_i P_j (m_i - m_j)(m_i - m_j)^T$$

The optimal features are determined by optimizing the Fisher criteria given by

$$J_{DAFE} = tr(S_{wY}^{-1} S_{bY}) \qquad (1)$$

Approximated pairwise accuracy criterion Linear Dimension Reduction (aPAC-LDR) [2] can be seen as DAFE weighted contributions of individual class pairs according to the Euclidian distance of respective class means. The major difference between DAFE and aPAC-LDR is that the Fisher criteria is redefined as

$$J_{a-PAC}(A) = \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} P_i P_j \omega(\Delta_{ij}) tr[(A S_{wX} A^T)^{-1}(A S_{ij} A^T)],$$

where $S_{ij} = (m_i - m_j)(m_i - m_j)^T$, $\Delta_{ij} = \sqrt{(m_i - m_j)^T S_w^{-1}(m_i - m_j)}$

and $\omega(\Delta_{ij}) = \dfrac{1}{2\Delta_{ij}} erf(\dfrac{\Delta_{ij}}{2\sqrt{2}})$

The above weighted Fisher criteria is the same as (1) by redefining the between-class scatter matrix as

$$S_b = \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} P_i P_j \omega(\Delta_{ij})(m_i - m_j)(m_i - m_j)^T$$

Hence the optimization problem is the same as DAFE.

The advantage of DAFE (or aPAC-LDR) is that it is distribution-free but there are three major disadvantages in DAFE (or aPAC-LDR). One is that it works well only if the distributions of classes are normal-like distributions [1]. When the distributions of classes are nonnormal-like or multi-modal mixture distributions, the performance of DAFE is not satisfactory. The second disadvantage of DAFE is the rank of the within-scatter matrix S_w is number of classes (L) –1, so generally only L-1 features can be extracted. In real situations, the data distributions are often complicated and not normal-like, therefore only using L-1 features is not sufficient for much real data. The third limitation is that if the within-class covariance is singular, which often occurs in high dimensional problems, DAFE will have a poor performance on classification.

Nonparametric Discriminant Analysis (NDA) [1][3] was proposed to solve the problems of DAFE. In NDA, the between-class scatter matrix is redefined as a new nonparametric between-class scatter matrix (for the 2 classes problem), S_b denoted, as

$$\begin{aligned} S_b &= P_1 E\{(X^{(1)} - M_2(X^{(1)}))(X^{(1)} - M_2(X^{(1)}))^T \mid \omega_1\} + \\ &\quad P_2 E\{(X^{(2)} - M_1(X^{(2)}))(X^{(2)} - M_1(X^{(2)}))^T \mid \omega_2\} \end{aligned} \qquad (2)$$

where $M_i(X) = \dfrac{1}{k}\sum_{j=1}^{k} x_{jNN}^{(i)}$ is called the local kNN mean of X, $x_{jNN}^{(i)}$ is the jth nearest neighborhood (NN) from class i (ω_i) to the sample x_l, and $x^{(i)}$ refers to samples from class i . The parametric S_w was still suggested to be used in NDA by the authors.

The disadvantages of NDA are: 1.Parameters k and □ are usually decided by rules of thumb. So the better result usually comes after several trails. 2. S_w is still with a parametric form. When the training set size is small, NDA will have the singularity problem.

II. Nonparametric Weighted Feature Extraction

Nonparametric weighted feature extraction (NWFE) [4] is proposed for improving DAFE and NDA In NWFE, the nonparametric between-class scatter matrix for L classes is defined as

$$S_b = \sum_{i=1}^{L} P_i \sum_{\substack{j=1 \\ j\neq i}}^{L} \sum_{k=1}^{n_i} \frac{\lambda_k^{(i,j)}}{n_i} (x_k^{(i)} - M_j(x_k^{(i)}))(x_k^{(i)} - M_j(x_k^{(i)}))^T \quad (3)$$

where $x_k^{(i)}$ refers to the k-th sample from class i. Basically, (3) is similar to (2). The differences are in the definitions of weights and local means. The scatter matrix weight $\lambda_k^{(i,j)}$ is a function of $x_k^{(i)}$ and $M_j(x_k^{(i)})$, and defined as:

$$\lambda_k^{(i,j)} = \frac{dist(x_k^{(i)}, M_j(x_k^{(i)}))^{-1}}{\sum_{l=1}^{n_i} dist(x_l^{(i)}, M_j(x_l^{(i)}))^{-1}}$$

where *dist(a,b)* means the distance from a to b. If the distance between $x_k^{(i)}$ and $M_j(x_k^{(i)})$ is small then its weight $\lambda_k^{(i,j)}$ will be close to 1; otherwise, $\lambda_k^{(i,j)}$ will be close to 0 and sum of total $\lambda_k^{(i,j)}$ for class *i* is 1. $M_j(x_k^{(i)})$ is the local mean of $x_k^{(i)}$ in the class j and defined as:

$$M_j(x_k^{(i)}) = \sum_{l=1}^{n_j} w_{kl}^{(i,j)} x_l^{(j)}$$

$$\text{where } w_{kl}^{(i,j)} = \frac{dist(x_k^{(i)}, x_l^{(j)})^{-1}}{\sum_{l=1}^{n_j} dist(x_k^{(i)}, x_l^{(j)})^{-1}} .$$

The weight $w_{kl}^{(i,j)}$ for computing local means is a function of $x_k^{(i)}$ and $x_k^{(j)}$. If the distance between $x_k^{(i)}$ and $M_j(x_k^{(i)})$ is small then its weight $\lambda_k^{(i,j)}$ will be close to 1; otherwise, $\lambda_k^{(i,j)}$ will be close to 0 and sum of total $\lambda_k^{(i,j)}$ for class *i* is 1. The nonparametric within-class scatter matrix is defined as

$$S_w = \sum_{i=1}^{L} P_i \sum_{k=1}^{n_i} \frac{\lambda_k^{(i,i)}}{n_i} (x_k^{(i)} - M_i(x_k^{(i)}))(x_k^{(i)} - M_i(x_k^{(i)}))^T$$

III. Experiment Design and Results

The simulated and real data set performances of five methods, DAFE, aPAC-LDR, NWFE, and NDA using 1NN and 5NN based on the □=2, were compared under experiment 1, 2, 3 and 4. Table 1 and 2 are the designs of simulated data experiment 1, 2, and 3 including training, testing samples and the dimensionality of data sets. The Washington, DC Mall image data was used for real data experiment. There are 7 classes in it and 40 training samples in each class.

Table 1 Experiment Design of Experiment 1 and 2

Dim=60		class 1	class 2	class 3	class 4	class 5	class 6
Mean		[0,…,0]	[1,0,…,0]	[0,1,0,…,0]	[0,0,1,0,…,0]	[1,1,0,…,0]	[1,0,1,0,…,0]
Cov	Exp1	0.1I					
	Exp2	0.1I		0.2I		0.3I	
Training		40	40	40	40	40	40
Testing		400	400	400	400	400	400

Table 2 Experiment Design of Experiment 3

Dim=60	class 1		class 2		class 3	
	comp 1	comp 2	comp 1	comp 2	comp 1	comp 2
Mean	[2,2,0,…,0]	[0,0,…,0]	[2,4,…,0]	[4,-2,0,…,0]	[-2,0,…,0]	[6,0,…,0]
Cov	0.1I					
Training	20	20	20	20	20	20
Testing	200	200	200	200	200	200
	class 4		class 5		class 6	
	comp 1	comp 2	comp 1	comp 2	comp 1	comp 2
Mean	[-2, 2,0,…,0]	[0,6,…,0]	[2,-4,…,0]	[-4,2,0,…,0]	[2,0,…,0]	[-6,0,…,0]
Cov	0.1I					
Traininge	20	20	20	20	20	20
Testing	200	200	200	200	200	200

Fig. 1, and 2 show the mean of accuracies of 10 simulated data sets for experiment 1, and 3. Fig. 3 show the mean of accuracies of 10 real data sets selected from the DC Mall image data with 191 bands. All figures show that NWFE performs better than the other methods in all experiments. For mixture distribution data, NWFE performs significantly better than the DAFE, aPAC-LDR, and NDA whether the dimensionality is large or not. Fig. 4 shows a color IR image of a portion of the DC Mall area for reference. Fig. 5, and 6 are the classified DC Mall maps for DAFE and NWFE respectively. Obviously, the result of NWFE is better than that of DAFE.

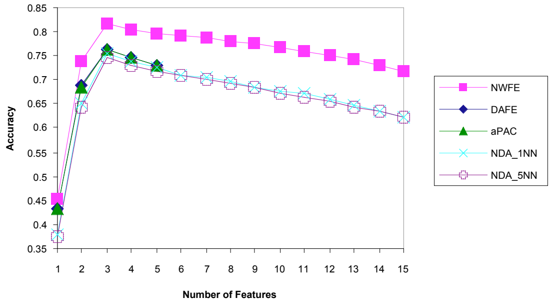

Fig. 1. Mean of accuracies of Experiment 1

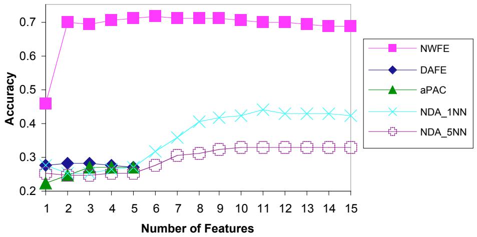

Fig. 2. Mean of accuracies of Experiment 3

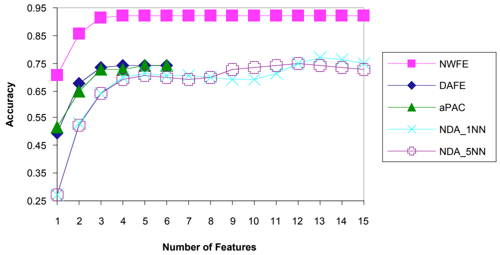

Fig. 3. Mean of accuracies of Experiment 4

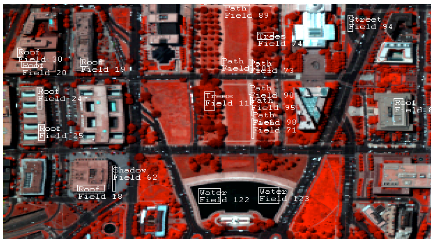

Fig. 4. A color IR image of a portion of the DC data set.

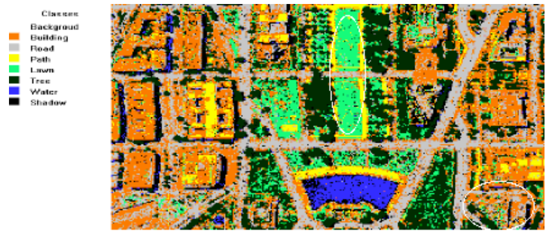

Fig. 5. The thematic map resulting from the classification of the area of Fig. 4 using DAFE features.

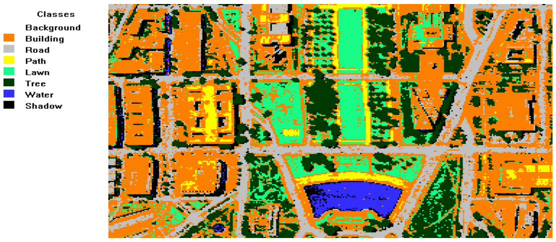

Fig. 6. The thematic map resulting from the classification of the area of Fig. 4 using NWFE features

IV. Concluding comments

The NWFE algorithm presented here is intended to take advantage of the desirable characteristics of DAFE, aPAC-LDR, and NDA, while avoiding their shortcomings. DAFE is fast and easy to apply, but its limitation of L-1 features, its reduced performance particularly when the difference in mean values of classes is small, and the fact that it is based on the statistical description of the entire training set, making it sensitive to outliers, limit its performance in many cases. NDA does not have these limitations and focuses the attention on training samples near the needed decision boundary. NDA does not perform well on unequal covariance or complexly distributed data.

NWFE does not have any of these limitations. It appears to have improved performance in a broad set of circumstances, making possible substantially better classification accuracy in the data sets tested, which included sets of agricultural, geological, ecological and urban significance. This improved performance is perhaps due to the fact that, like NDA, attention is focused upon training samples that are near to the eventual decision boundary, rather than equally weighted on all training pixels as with DAFE. It also appears to provide feature sets which are relatively insensitive to the precise choice of feature set size, since the accuracy versus dimensionality curves are relatively flat beyond the initial knee of the curve. This characteristic would appear to be significant for the circumstance when this technology begins to be used by general remote sensing practitioners who are not otherwise highly versed in signal processing principles and thus might not realize how to choose the right dimensionality to use.

V. Reference

[1] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, San Diego: Academic Press Inc., 1990

[2] R, P. W. Duin and R. Haeb-Umbach, “Multiclass Linear Dimension Reduction by Weighted Pairwise Fisher Criteria,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, 2001, pp. 762-766.

[3] K. Fukunaga and M. Mantock, Nonparametric Discriminant Analysis, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 5, 1983, pp. 671-678.

[4] Bor-Chen Kuo and David Landgrebe, Improved Statistics Estimation And Feature Extraction For Hyperspectral Data Classification, PhD Thesis and School of Electrical & Computer Engineering Technical Report TR-ECE 01-6, December 2001 (88 pages).