LINEAR FEATURE EXTRACTION FOR MULTICLASS PROBLEMS

Pi-Fuei Hsieh and David Landgrebe
School of Electrical & Computer Engineering
Purdue University
West Lafayette IN 47907-1285
hsieh@ecn.purdue.edu

ABSTRACT

Motivated by the need for a fast and effective feature extraction method for multiclass problems, a feature extraction method is developed to satisfy two requirements: (1) perform on a class-statistics basis (2) use discriminant information about covariance-difference as well as mean-difference. Experiments show that the new feature extraction method has fulfilled the requirements when the number of training samples is large. Experiments with a small number of training samples were also conducted for showing the limitation of feature extraction.

INTRODUCTION

In analyzing hyperspectral data, the information about discriminating among classes is quite often contained primarily in a smaller number of features than the number of measurements (in channels). In order to make classification effective and efficient, it is desirable to extract these informative features. Feature extraction can be considered as a mapping from the original dimensional space to a lower dimensional space, where class separability is approximately preserved. Since it is difficult to perform nonlinear transformations, the discussion in this study will be limited to linear feature extraction. In parametric feature extraction, there are two types of discriminatory information: mean-difference and covariance-difference. For general purpose use, a feature extraction method combining mean-difference and covariance-difference is desired.

For multiclass problems, Discriminant Analysis Feature Extraction (DAFE) [1][2] and Decision Boundary Feature Extraction (DBFE) [3] are two reliable schemes currently being used. DAFE is fast and easy to implement. However, the application of DAFE is limited to the case in which classes have significant mean-difference since DAFE is based upon the information about mean-difference. The number of effective features extracted by DAFE is subject to the number of classes. Although DBFE does not have these disadvantages, DBFE is very time-consuming. It is a numerical algorithm performing on a sample-by-sample basis. The time complexity of DBFE is a function of the number of training samples. A good performance of DBFE usually demands a large number of training samples. In addition, an extension to the Foley-Sammon orthonormal feature extraction method has been proposed [4][5]. However, this extension was not effective (see Experiments).

Motivated by the need for a fast and effective feature extraction method for multiclass problems, a feature extraction method is developed that satisfies the following two requirements. First, this method should perform on a class-statistics basis so that it is faster than DBFE. Second, this method should use discriminant information about covariance-difference as well as mean-difference so that it may generate more effective features than DAFE.

In order to utilize the discriminant information about covariance-difference as well as mean-difference, a two-stage strategy is used. This strategy has been proposed for two-class problems [6]. The idea behind this strategy is as follows. In the nullspace of the row DAFE features, classes have common mean vectors. That is, DAFE is followed by a common-mean problem. Bounds on the classification accuracy for common-mean problems are derived for the criterion used at the second stage.

It is shown that the upper and lower bounds are associated with the ratio of the largest to the smallest class variances. This leads to the conclusion that the most effective feature can be selected by picking the feature along which the ratio of the largest to the smallest variances is highest.

Experiments show that the new feature extraction method has fulfilled the requirements. When the number of training samples is large, the proposed method extracts more effective features than DAFE and needs much less computational time than DBFE while having a comparable performance to DBFE.

BOUNDS ON THE CLASSIFICATION ACCURACY

The classification along each feature can be considered as a univariate problem. In the common-mean multiclass case, it is easier to formulate the classification accuracy than to formulate the Bayes error; thus, a mathematical expression was derived for the classification accuracy [7]. Since the expression, involved with an integration, is not a closed form, bounds are provided to gain insight into the discriminant information.

Theorem 1: For L univariate normal distributions with a common mean and different variances, $\sigma_1^2 < \sigma_2^2 < \ldots < \sigma_L^2$, the probability of correct classification under the maximum likelihood classifier can be bounded as follows.

$$(1/L)\left[1 + 2\,\Phi\left(d_{1L}/\sigma_1\right) - 2\,\Phi\left(d_{1L}/\sigma_L\right)\right] \le$$
$$P_{cr} \le (1/L)\left[1 + (2\pi e)^{-1/2}\,\ln\left(\sigma_L^2/\sigma_1^2\right)\right]$$

where P_{cr} = classification accuracy; d_{ij} = the decision point between class i and class j on the side of $x>0$, $d_{ij}^2 = [\sigma_i^2\sigma_j^2 / (\sigma_j^2 - \sigma_i^2)]\ln(\sigma_j^2 / \sigma_i^2)$; $\Phi(x)$ is the cumulative distribution function of the standard normal distribution, $\Phi(x) = (2\pi)^{-1/2}\int_{-\infty}^{x}\exp(-t^2/2)dt$.

Theorem 2: Assume that the covariance matrices of L equally-probable classes share the same eigenvectors, then the overall classification accuracy (P_c) is bounded by

$$P_c \le L^{n-1}\prod_{k=1}^{n} P_{cr,k}$$

where $P_{cr,k}$ is the classification accuracy along the k-th eigenvector.

This theorem suggests that the best m-dimensional subspace be spanned by the m features corresponding to the m highest accuracy rates if classes share the same eigenvectors. It is shown in Theorem 1 that the larger the ratio of variances, the higher the classification accuracy along the feature. Therefore, it can be concluded that the best m features can be selected from the eigenvectors corresponding to the m largest variance ratios.. When $L=2$, this conclusion is the same as [1].

CRITERIA FOR FEATURE EXTRACTION

Criterion I: A linearly independent feature set

Given a feature φ, the ratio of the largest to the smallest variance is $\rho_\varphi = \max_{i,j,i\neq j}\frac{\varphi^T\Sigma_i\varphi}{\varphi^T\Sigma_j\varphi}$, where Σ_i is the covariance matrix of class i. The best feature can be obtained by maximizing ρ_φ over all possible φ. Thus, a criterion for a linearly independent feature set is:

$$\max_{\varphi}\max_{i,j,i\neq j}\frac{\varphi^T\Sigma_i\varphi}{\varphi^T\Sigma_j\varphi} = \max_{i,j,i\neq j}\max_{\varphi}\frac{\varphi^T\Sigma_i\varphi}{\varphi^T\Sigma_j\varphi}.$$

The optimization of $\frac{\varphi^T\Sigma_i\varphi}{\varphi^T\Sigma_j\varphi}$ for given i and j is equivalent to the optimization of $tr(\Sigma_j^{-1}\Sigma_i)$., and $\lambda_{max}^{(ij)}$ is the largest eigenvalue of $\Sigma_j^{-1}\Sigma_i$. The corresponding eigenvector $\varphi_{max}^{(ij)}$ of $\Sigma_j^{-1}\Sigma_i$ maximizes the criterion for discriminating class i and class j.

Algorithm:

1. For each pair of class i and class j ($i \neq j$), compute the eigenvalues and eigenvectors of $\Sigma_j^{-1}\Sigma_i$.
2. Sort all eigenvalues in a decreasing order. The eigenvector corresponding to the largest eigenvalue is the best feature. Put it in the feature set.
3. Check the next largest eigenvalue. Retain the corresponding eigenvector if it is linearly independent of the existing feature set. Add the vector to the feature set. Repeat this step until n features are obtained.

Two other criteria were also developed for test.

Criterion II: An orthogonal feature set:

$$\max_{i,j,i\neq j}\max_{\varphi}\frac{\varphi^T\Sigma_i\varphi}{\varphi^T\Sigma_j\varphi}$$

subject to $\varphi_h^T\varphi = 0,\ h = 1,\ldots,k-1$

Criterion III: A linearly independent feature set 2

$$\max_{i,j,i\neq j}\max_{\Phi}\frac{|\Phi^T\Sigma_i\Phi|}{|\Phi^T\Sigma_j\Phi|}$$

where Φ is an n by p matrix consisting of p best linearly independent features regarding the pair of classes i and j. p is specified by users.

EXPERIMENTS

The real data used in the experiment was gathered by Spectrometer System (FSS), a helicopter-mounted filed spectrometer. This data set consisted of four classes that were chosen from the data collected at Finney Co. KS. on May 3 and March 8, 1977 (Table 1). To guarantee that the ratio of training sample size to dimensionality was large enough, a simple dimension reduction was conducted. The number of dimensions was reduced from 60 to 20 by combining every three consecutive bands. The following cases were considered.

Test-1: Common-mean case with large training sizes
Test-2: Different-mean case with large training sizes
Test-3: Common-mean case with small training sizes
Test-4: Different-mean case with small training sizes

Training samples were randomly selected from the corresponding real data set, and the rest of the samples were all used as test samples. A size of 300 was used for the large training size case whereas 22 was used for the small training size case.

Common-mean data sets were prepared by projecting the 20-dimensional data to the 17-dimensional subspace orthogonal to the subspace spanned by the DAFE features, with respect to the within-class scatter matrix S_w. At this step, the class statistics were estimated (referred to as the true statistics) using all samples; discriminant features were extracted by DAFE; and data were transformed to the nullspace of the row DAFE features. The resulting classes had common mean vectors.

For different-mean cases, the feature extraction at the second stage can be performed in either of the following two subspaces. One is orthogonal to the DAFE subspace with respect to the within-class scatter matrix S_w, and the other is orthogonal to the DAFE subspace. The latter was used in the experiments.

The criteria I, II, and III and other methods were tested, including DBFE, DAFE, an extension to Foley-Sammon, and DAFE-DAFEs. DAFE-DAFEs can be considered as another extension to Foley-Sammon that selects L-1 features rather than one feature at a time.

Experimental results are shown in Figures 1-4. Several observations were made. First, when the number of training samples was large, the proposed criteria extracted more effective features than DAFE and needed much less

computational time than DBFE while having a comparable performance to DBFE. This implies that incorporating the information about covariance-difference into feature extraction helps improve the performance. Second, the extension to Foley-Sammon gave poor results, indicating that the orthogonality constraint is not necessarily helpful and the way to extract features one by one is not appropriate. Third, in the case of small training sample sizes, the performance of feature extraction methods was contrary to the order in the large training case. DBFE which used the information about covariance-difference gave the poorest performance among all methods. This is reasonable because the estimates of covariances are no longer reliable when the number of training samples is small. Inaccurate estimation of class statistics undermines the performance of feature extraction.

Table 1: The FSS1977 Data Set

Class Name	No. of samples
1. Winter Wheat, May 3	657
2. Unknown Crops, May 3	678
3. Winter Wheat, March 8	691
4. Unknown Crops, March 8	619

CONCLUSIONS

Linear parametric feature extraction for multiclass problems has been investigated. The objective is to develop a fast and effective feature extraction method that utilizes discriminant information about covariance-difference as well as mean-difference and performs on a class-statistics basis. The proposed two-stage feature extraction method has achieved the goal in the experiments that have been done. It should be noted that such feature extraction methods are not suitable for the case of small training sample sizes.

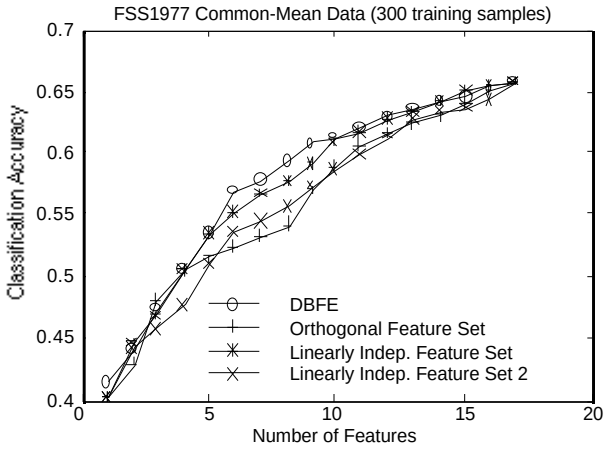

Figure 1: Common-mean case with large training sizes (DBFE: Decision Boundary Feature Extraction)

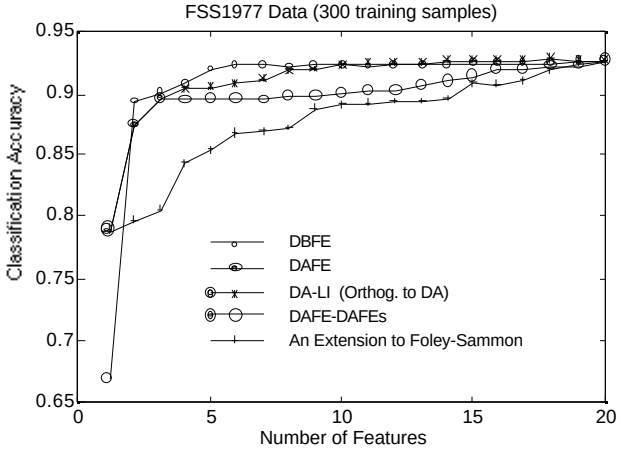

Figure 2: Different-mean case with large training sizes. (DA-LI denotes DAFE followed by the linear independent feature extraction based on Criterion I.)

REFERENCES

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

[2] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugen*., vol. 7, pp. 179-188, 1936.

[3] C. Lee and D. A. Landgrebe, "Feature extraction based on decision boundaries," IEEE Trans. Pattern Anal. Machine Intell., vol. 15, no. 4, pp. 321-325, April 1993.

[4] T. Okada and S. Tomita, "An optimal orthonormal system for discriminant analysis," *Pattern Recognition*, vol. 18, no. 2, pp. 139-144, 1985.

[5] J. Duchene and S. Leclercq, "An optimal transformation for discriminant and principal component analysis", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-10, no. 6, pp. 978-983, Nov. 1988.

[6] I. D. Longstaff, "On extensions to Fisher's linear discriminant function," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-9, no. 2, pp. 321-325, March 1987.

[7] P. Hsieh, "Classification of High Dimensional Data," TR-ECE, 98-04, Purdue University, May 1998.

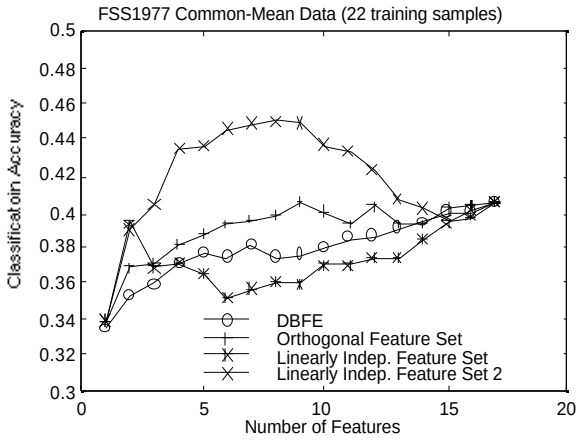

Figure 3: Common-mean case with small training sizes

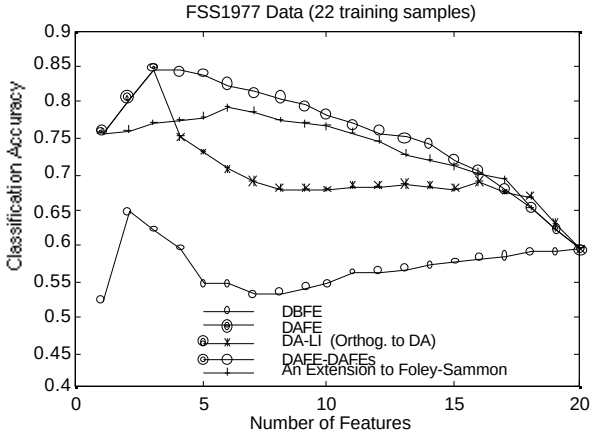

Figure 4: Different-mean case with small training sizes