STATISTICS ENHANCEMENT IN HYPERSPECTRAL DATA ANALYSIS USING SPECTRAL-SPATIAL LABELING, THE EM ALGORITHM, AND THE LEAVE-ONE-OUT COVARIANCE ESTIMATOR

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

ABSTRACT

Hyperspectral data potentially contain more information than multispectral data because of higher dimensionality. Information extraction algorithm performance is strongly related to the quantitative precision with which the desired classes are defined, a characteristic which increases rapidly with dimensionality. Due to the limited number of training samples used in defining classes, the information extraction of hyperspectral data may not perform as well as needed. In this paper, schemes for statistics enhancement are investigated for alleviating this problem. Previous works including the EM algorithm and the Leave-One-Out covariance estimator are discussed. The HALF covariance estimator is proposed for two-class problems by using the symmetry property of the normal distribution. A spectral-spatial labeling scheme is proposed to increase the training sample sizes automatically. We also seek to combine previous works with the proposed methods so as to take full advantage of statistics enhancement. Using these techniques, improvement in classification accuracy has been observed.

Keywords: Hyperspectral analysis, multispectral analysis, classification, the EM algorithm, parameter estimation, small sample size problem.

1. INTRODUCTION

In the case of hyperspectral data, it has been found that adequate training for classifiers is of even greater importance than in the case with conventional multispectral data. Adequate training depends on accurate parameter estimation, where a large number of training samples, compared to the dimensionality, is usually required. When the number of training samples is relatively small, maximum likelihood estimates of the parameters have a large amount of variance, leading to a large classification error [1]. Furthermore, as second order statistics deserve attention no less than first order statistics in hyperspectral analysis [2], there is an even greater demand for accurate parameter estimation.

In general, classification errors depend on the accuracy of statistics estimation, the classifier type, and the class separability. One can seek to improve the classification performance from any of these aspects [3]. In this paper, the factor of statistics estimation is singled out for consideration. Previous works, including the EM algorithm [4][5] and the Leave-One-Out covariance estimator [6], are reviewed briefly. Two new methods are proposed. A training sample labeling method based on spectral and spatial information is proposed for automatically gathering a larger number of training samples. A fast method based on the symmetry property of the normal distributions is proposed specially for two-class problems. The combinations of these methods are also sought in order to achieve the best performance of statistics enhancement. Experimental results are given and discussed in Section 5.

2. PREVIOUS WORKS

2.1. The Expectation Maximization (EM) algorithm

Use of the EM algorithm, recently proposed as a means of supervised-unsupervised learning [5], incorporates unlabeled samples into the process of parameter estimation. The goal of the EM algorithm is to find the optimal choice of statistics that maximizes the log-likelihood function L given by

Presented at the SPIE International Symposium on Optical Science, Engineering, and Instrumentation, San Diego, CA, July 19-24, 1988

$$L = \sum_{k=1}^{K} log\, f(x_k) + \sum_{i=1}^{L} \sum_{k=1}^{N_i} log\left[P_i p_i (z_{ik} | \mu_i, \Sigma_i) \right]$$

where $p_i(x_k|\mu_i, \Sigma_i)$ is the conditional density function of class i, i=1..L, and $f(x_k)$ is the mixture density given by

$$f(x_k) = \sum_{i=1}^{L} P_i p_i (x_k | \mu_i, \Sigma_i)$$

where z and x denote training and unlabeled samples, respectively.

Using unlabeled samples has an equivalent effect of increasing training sample size, though unlabeled samples may not be as effective as training samples. In practice, the performance of the EM algorithm is not always reliable. The EM algorithm may converge to a local maximum of the likelihood function, resulting in poor estimates of class statistics. Precaution should be exercised as to whether or not the classes are normally distributed and spectrally separable[5][3], the list of classes is exhaustive, and the initial statistics are reliable. In this paper, enhancing initial statistics is considered for improving the performance.

2.2. The Leave-One-Out Covariance (LOOC) estimator

The Leave-One-Out Covariance estimator (LOOC) [6] is similar to the Regularized Discriminant Analysis (RDA) [7]. By examining pair-wise linear combinations of diagonal covariance, covariance, within-class covariance, and diagonal within-class covariance matrices, the LOOC method determines the best combination that maximizes the following average log likelihood function of leave-one-out test samples:

$$L_i(\alpha) = \frac{1}{N_i} \sum_{k=1}^{N_i} ln p(x_k | \mu_{i/k}, C_{i/k}(\alpha)),$$

where $\ln p(x_k | \mu_{i/k}, C_{i/k}(\alpha))$ is the log likelihood value of the sample x_k of class i. The mean $\mu_{i/k}$ and the mixed covariance $C_{i/k}(\alpha)$ are computed without sample x_k, where α is the mixing parameter of the covariance combination. Such combining approach may or may not outperform the traditional ML estimates (or sample estimates). However, in the case of small training sample sizes, this approach has shown the advantages that it improves the classification performance and provides a nonsingular covariance matrix (if there are at least three independent samples in a class and these samples do not have the same value on any band).

3. NEW SCHEMES

3.1. HALF

This is a fast covariance estimator for two-class problems based on the property that a normal distribution is symmetric with respect to the class mean. This symmetry property can be used to estimate the covariance matrix when the samples in one half space are contaminated, missing, or mixed with other distributions. The covariance matrix can be estimated by using the samples on the other half space with respect to the class mean. If two points, x and y, satisfies $(x + y)/2 = \mu$, where μ is the class mean, then the probability densities $p(x|\mu, \Sigma) = p(y|\mu, \Sigma)$, and $(x-\mu)(x-\mu)^T = (y-\mu)(y-\mu)^T$. Consider a normal distribution $N(\mu, \Sigma)$ where μ is known and Σ is unknown. Let v be the norm vector of a hyperplane passing the class mean. Let Ω be the sample set containing K samples in one half space in the direction of v: $\Omega = \{X | (X-\mu)^T v ? 0\}$, where X denotes a sample. The covariance matrix can be estimated from Ω by $\tilde{\Sigma} = (1/K) \sum (X_i - \mu)(X_i - \mu)^T, X_i \in \Omega$. $\tilde{\Sigma}$ is an unbiased estimate of the covariance matrix. For a univariate normal distribution, $N(0, \sigma^2)$, $VAR(\tilde{\sigma}^2) = 2\sigma^4 / K$, which is approximately two times larger than that of the sample variance estimate. The algorithm for two-class problems is described as follows:

1. Compute sample mean vectors $\hat{\mu}_1, \hat{\mu}_2$ for each class by using training samples.

2. Let $v = \hat{\mu}_2 - \hat{\mu}_1$ and unlabeled sample sets Ω_1 and Ω_2 be

$$\Omega_1 = \left\{ x_k \middle| (x_k - \hat{\mu}_1)^T v \le 0 \right\}$$
$$\Omega_2 = \left\{ x_k \middle| (x_k - \hat{\mu}_2)^T v \; ? \; 0 \right\}$$

3. Estimate class covariance matrices using training samples and unlabeled samples in Ω_1 and Ω_2 for class 1 and class 2, respectively:

$$\hat{\Sigma}_i = \frac{\sum_{k=1}^{N_i} (z_{ik} - \hat{\mu}_i)(z_{ik} - \hat{\mu}_i)^T + \sum_{x_k \in \Omega_i} (x_k - \hat{\mu}_i)(x_k - \hat{\mu}_i)^T}{N_i + K_i}$$

where

z_{ik} = the training samples of class i, $k = 1,.., N_i$;
x_k = unlabeled samples;
N_i = the number of training samples from class i;
K_i = the size of Ω_i.

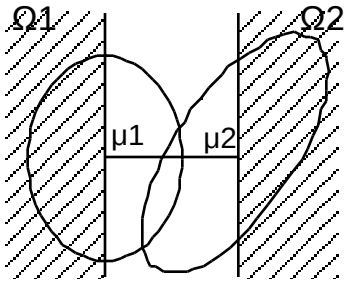


Fig. 1. For two classes, the covariance matrices can be estimated by using the samples in the shaded regions.

The HALF method is a fast estimation method that is designed especially for two-class problems. Based on the symmetry property of normal distributions, unlabeled samples are selected to enhance the initial estimation of covariance matrices. Due to the nature of the symmetry property, only two-class problems are considered. If the class mean is unknown, the performance of this covariance estimator will depend highly on the accuracy of the class mean estimation. Experimental results are given in section 5.1.

3.2. Spectral-Spatial Training Sample Labeling

When the number of training samples is too small to provide sufficient information for defining classes, auxiliary information is helpful. A remotely sensed data set often consists of regions, each containing one class of ground cover type, therefore, pixels of the same class are likely to be spatially contiguous. Besides, a data set often abounds in unlabeled samples that may belong to classes of interest. Noting these two inherent characteristics, we have developed the spectral-spatial training-sample labeling method. A previous work [7] related to training sample labeling is based on the fact that many minerals have unique and diagnostic absorption characteristics in their reflectance spectra. Training samples in mineral classification problems can be labeled if laboratory reflectance spectra are available. The spectral-spatial labeling method proposed here utilizes spatial contiguity and the characteristics of normal distributions in the feature space.

Algorithm:

1. Compute the sample mean ($\hat{\mu}_i$) and the sample covariance for each class by using the training samples.

2. If a sample covariance is singular, pick a covariance ($\hat{\Sigma}_i$) among the following choices.
   a. The diagonal of the sample covariance. (That is, assume that the bands are uncorrelated.)

Presented at the SPIE International Symposium on Optical Science, Engineering, and Instrumentation, San Diego, CA, July 19-24, 1988

b. The covariance matrix of the entire data set.
c. The average of the sample covariance matrices over all classes.

3. In the feature space, label a sample to the nearest class if the sample falls inside a given probability region of the nearest class. The nearest class j is determined by using the maximum likelihood criterion given by

$$j = \arg\min_i (X - \hat{\mu}_i)^T \hat{\Sigma}_i^{-1} (X - \hat{\mu}_i) + \ln\left|\hat{\Sigma}_i\right| .$$

A sample X falls inside a probability region with a probability mass of α if

$$(X - \hat{\mu}_j)^T \hat{\Sigma}_j^{-1} (X - \hat{\mu}_j) \le d_\alpha ,$$

where d_α is determined by $\Pr(\mathbf{d} \le d_\alpha) = \alpha$ for a given α, and $\mathbf{d}$ possesses a χ^2 distribution with n degrees of freedom (n = dimensionality). d_α remains constant for all classes. Repeat this step for each sample.

4. In the spatial domain, adjust labels as follows.

(i) Consider each labeled sample. If the majority of the neighboring samples belong to another class k, change the class ownership of this labeled sample to k if $(X - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1} (X - \hat{\mu}_k) \le d_\alpha$. If $(X - \hat{\mu}_k)^T \hat{\Sigma}_k^{-1} (X - \hat{\mu}_k) > d_\alpha$, remove the label so that the labeled sample becomes unlabeled. If there is no majority class, leave the labeled sample as it is. The neighborhood selected in this paper is a 3x3 window centered by the sample being considered.

(ii) Consider each unlabeled sample. If the majority of the neighboring labeled samples belong to class k, assign the unlabeled sample to k if the local homogeneity test is passed. The test for local homogeneity was within three times standard deviation distance from the class mean for each channel.

5. Update mean and covariance for each class using labeled samples. Repeat Step 3 through Step 5 several times until a satisfactory result is obtained.

The intermediate result at each step is generated in order to keep track of the change. Special care has been taken in step 4, where labels are to be adjusted. Note that steps 4(i) and 4(ii) share the same criterion for examining spatial context, but they are separated into two steps. The reason for this design is to avoid carrying the error due to mislabeled samples into the labeling of unlabeled samples. Furthermore, a homogeneity test is added at the step 4(ii) to impose a stricter condition on the labeling of unlabeled samples.

This labeling method can be used to automatically label likely training samples and augment the size of the training sample set selected by a human analyst. The method utilizes the information in the feature space and in the spatial space. However, based primarily on local spatial information, this method does not have a global optimization criterion in support of the labeling scheme.

4. STATISTICS ENHANCEMENT

Because each method has its own weakness and benefit, it is desirable to consider their combinations so as to supplement each other. The LOOC estimator can provide nonsingular covariance matrices as the initial setting for the EM algorithm and the spectral-spatial labeling method. The EM algorithm can provide a global optimization criterion in support of the labeling scheme. The LOOC estimator uses only the information about training samples whereas the EM algorithm and the labeling method gather unlabeled samples as well as training samples. Thus, for multiclass problems, four combinations are considered in this paper: 1. LOOC and Labeling; 2. LOOC and the EM algorithm; 3. Labeling and the EM algorithm; 4. LOOC, Labeling, and the EM algorithm.

5. EXPERIMENTS

5.1. Two-class Problems

Tests were performed on a computer-generated 8-dimensional data set. Assume that two classes possess normal distributions, N(0,I) and N([0.965 0.775 0.21 0.21 0.410 0.270 0.065 0.0025]^T, Σ_2), where Σ_2 is a diagonal

Presented at the SPIE International Symposium on Optical Science, Engineering, and Instrumentation, San Diego, CA, July 19-24, 1988

matrix whose diagonal elements are [8.41 12.06 0.12 0.22 1.49 1.77 0.35 2.73]). The Bayes error is about 0.8%. Eight training samples per class were generated for each trial. Test and unlabeled samples were fixed throughout all trials. There were 500 test samples for each class and 1000 unlabeled samples (500 from each class). 10 trials have been conducted. The EM algorithm was set to a maximum of 20 iterations. The quadratic Maximum Likelihood classifier was used, and the prior probabilities were assumed to be equal.

Note that the sample covariance matrices are singular matrices, which cannot be used for the EM algorithm or the quadratic classifier. To obtain nonsingular covariance matrices, the Nearest Mean [1], the LOOC, and the HALF methods were considered. The Nearest Mean and the HALF incorporate unlabeled samples. Since a reliable convergence of the EM algorithm depends on a good starting point, an initial statistics setting seems necessary for the EM algorithm in this case. The classification results are shown in Table 1. The LOOC outperforms others. If these nonsingular covariance matrices were used as the initial settings of the EM algorithm, all achieved the same classification accuracy of 91.23%.

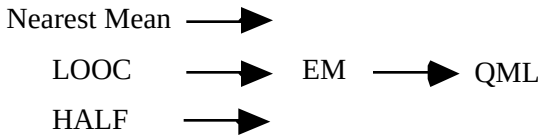


Fig. 2. Various procedures in the case of singular sample covariance matrices

Table 1. Classification accuracy of different covariance estimators

	Classification accuracy without using the EM Algorithm	Classification accuracy if followed by the EM Algorithm
Nearest Mean	60.75 %	91.23 %
LOOC	85.77 %	91.23 %
HALF	67.09 %	91.23 %

5.2. Multiclass Problems

For multiclass problems, experiments were conducted with an AVIRIS data set taken on July, 15, 1995 over the Indian Pine Test Site 1C (Fig. 3). From the ground truth, six classes were known: Corn, Corn-N, Soybean-NS15 (15" apart), Soybean-Drilled (7" apart), Soybean (30" apart), and Wheat. Due to different planting dates and practices, the degrees of the ground cover in corn or soybean fields varied even though fields were labeled to the same class in the ground truth map. From the spectral standpoint, it seems necessary to re-define classes based on the variety of ground cover levels. The classes of the ground truth were thus altered to be Corn-LowGroundCover, Corn-HighGroundCover, Soybean-LowGroundCover, Soybean-MediumGroundCover, Soybean-HighGroundCover, and Wheat. There were 36 training samples drawn from each class as shown in Fig. 3. The strategy for analysis is described as follows.

Removal of noisy channels: First of all, 39 channels were found noisy by viewing the channel-by-channel images. Most of them are water absorption channels. Retaining noisy channels may require an extra number of training samples so as to maintain classification performance [8]. Therefore, it is important to discard these redundant channels when there are only a limited number of training samples available.

Reduction of dimensionality: After discarding noisy channels, there were 185 channels left. The dimensionality of 185 was still rather large compared to the training sample size of 36, so the dimensionality was further reduced by picking every sixth channel. The resulting 30 dimensions was close to the number of training samples. Alternative ways to reduce dimensionality include the Projection Pursuit Dimension Reduction[9], and Uniform Feature Design.

An exhaustive set of classes: It is desirable to define an exhaustive set of classes for an accurate classification and a reliable performance of the EM algorithm. An exhaustive list of classes can be obtained by defining new classes and by removing outliers. To define new classes, clustering algorithms and the classification probability map had been often considered. However, both of them did not work in this case due to a small number of training samples. To remove outliers, the most commonly used technique was to set a threshold for the squared distance between sample

X and the estimated mean vector $\hat{\mu}_i$ based on the chi-square distribution.

$$\Pr\{ (X - \hat{\mu}_i)^T \hat{\Sigma}_i^{-1} (X - \hat{\mu}_i) ? \eta_\alpha \} = \alpha$$

where α is a given probability (e.g. α =0.01). Note that the assumption that $(X - \hat{\mu}_i)^T \hat{\Sigma}_i^{-1} (X - \hat{\mu}_i)$ is a chi-square distribution is no longer appropriate if the number of training samples is small. Setting an adequate threshold becomes difficult.

Due to the difficulty of finding an exhaustive set of classes by using the above techniques, the spectral-spatial labeling scheme was used instead. In processing of the labeling, a blank area right above a soybean field suggested a tentative class. A new class was thus defined and referred to as 'Unknown Green'. Another blank area did not appear homogeneous, so no action was taken. Thus far, the list of classes has been finalized and a total of seven classes have been defined as seen in Table 2.

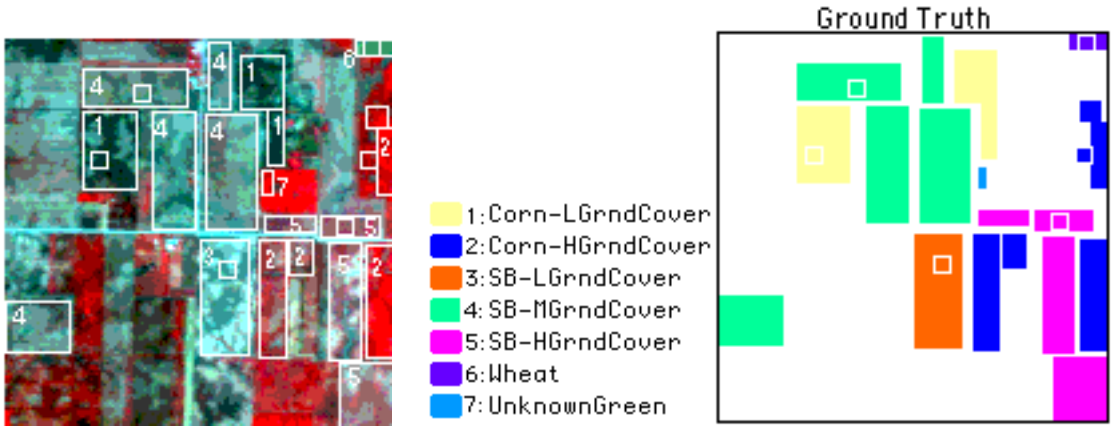
Schemes for Statistics Enhancement: Three schemes for statistics enhancement are considered: the spectral-spatial labeling method, the Leave-One-Out Covariance estimator (LOOC), and the EM algorithm. The spectral-spatial training-sample labeling method gathers likely training samples. The LOOC method gives the best linear combination of the pooled covariance, the original covariance and their diagonal covariance matrices. And the EM algorithm utilizes unlabeled samples. Because each has its own weakness and benefit, their combinations are considered so as to supplement each other. For instance, the performance of the EM algorithm depends on the initial statistics, so LOOC and the spectral-spatial labeling method may provide more reliable initial statistics for the EM algorithm. Since the spectral-spatial labeling method is based primarily on local spatial information, the subsequent use of the EM algorithm may provide a global optimization criterion in support of the labeling scheme. Therefore, the EM algorithm and the labeling scheme supplement each other. Through this experiment, the EM algorithm removed outliers based on the initial statistics before it started. The numbers of unlabeled samples used for the EM algorithm were 18968, 18477, 17847, and 18419 in EM, LOOC-EM, LOOC-Labeling-EM, and Labeling-EM, respectively.

Results: The experimental results are shown in Table 3 in order of classification performance. Fig. 4 shows the resulting classification maps from individual schemes and combination of schemes, respectively. When no statistics enhancement scheme was used, the classification performance was rather poor (66.6%). If any of the statistics enhancement schemes was used, the classification performance improved to be higher than 80%. When the schemes were combined, the classification performance improved further to be higher than 90%. The best performances were given by the combination of the Labeling method and the EM algorithm (95.01%) and the combination of all schemes (94.01%).

Table 2. Data Set: AVIRIS 1995 Indian Pine Test Site 1C

No. of classes	7	
No. of channels (original)	220	
No. of channels (reduced)	30	
Image size (in pixels)	145 x 145	
Total no. of samples	21025	
Total no. of training samples	252	
Total no. of test samples	7096	
No. of iterations of EM	10	
Class Name	Training Samples	Test Samples
1. Corn-LowGroundCover	36	1026
2. Corn-HighGroundCover	36	1191
3. Soybean-LowGroundCover	36	774
4. Soybean-MediumGroundCover	36	2723
5. Soybean-HighGroundCover	36	1298
6. Wheat	36	84
7. Unknown Green	36	0

Presented at the SPIE International Symposium on Optical Science, Engineering, and Instrumentation, San Diego, CA, July 19-24, 1988



(a) AVIRIS 1995 Indian Pine Test Site 1C (b)Ground truth map (Original in color)

Fig. 3. The AVIRIS data set taken on July 15, 1995 at Indian Pine Site 1C.

Presented at the SPIE International Symposium on Optical Science, Engineering, and Instrumentation, San Diego, CA, July 19-24, 1988

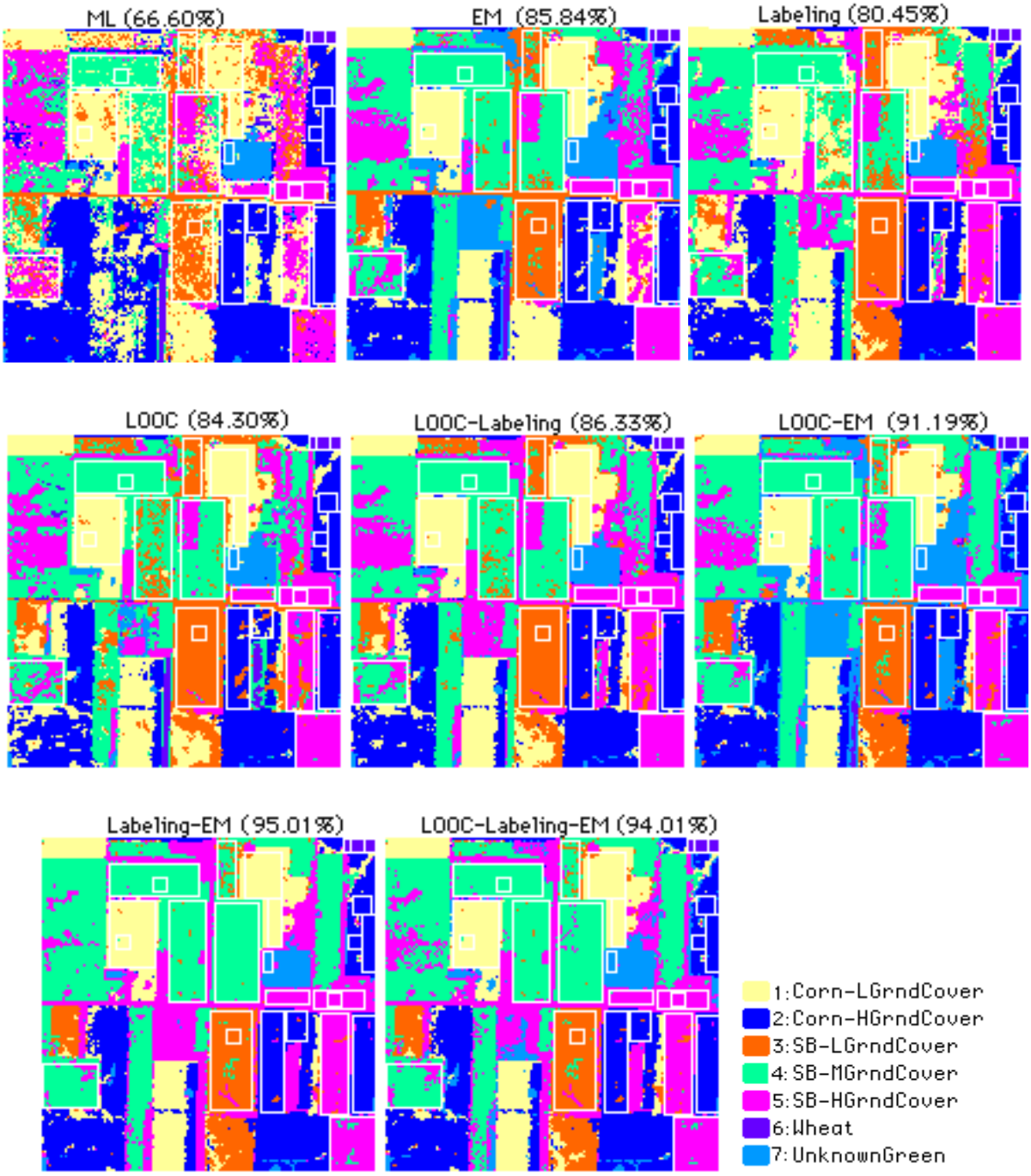


Fig. 4. Classification results of various procedures without and with statistics enhancement methods.

(Original in color)

Presented at the SPIE International Symposium on Optical Science, Engineering, and Instrumentation, San Diego, CA, July 19-24, 1988

Table 3. Comparison of various analysis procedures

Rank	Use of Statistics Enhancement	Statistics Enhancement Analysis Procedure			Accuracy (%)
8	(None)				66.60
7	Individual method		Labeling		80.45
6		LOOC			84.30
5				EM	85.84
4	Combination	LOOC	Labeling		86.33
3		LOOC		EM	91.19
2		LOOC	Labeling	EM	94.01
1			Labeling	EM	95.01

CONCLUSIONS

Statistics enhancement was considered in this paper. A spatial-spectral training sample labeling method has been developed in order to gather likely training samples and remove outliers. This new method was compared to another two methods: the EM algorithm and Leave-One-Out Covariance estimator (LOOC). Also, possible combinations of these three methods were investigated. It has been seen that a preprocessor of statistics enhancement is helpful for hyperspectral classification when the number of training samples is small. And an adequate combination of the methods can be used to take full advantage of each method, leading to further improvement.

ACKNOWLEDGMENT

Work reported in this paper was supported in part by NASA Grant NAG5-3975. This support is gratefully acknowledged.

Presented at the SPIE International Symposium on Optical Science, Engineering, and Instrumentation, San Diego, CA, July 19-24, 1988

REFERENCES

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

[2] C. Lee and D. A. Landgrebe, "Analyzing high dimensional multispectral data," *IEEE Trans. on Geosci. Remote Sensing*, vol. 31, no. 4, pp. 792-800, July 1993.

[3] P. Hsieh and D. A. Landgrebe, "Classification of high dimensional data," TR-EE 98-4, Purdue University, May 1998, available from http://dynamo.ecn.purdue.edu/~landgreb/publications.html.

[4] A. P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood estimation from incomplete data via EM algorithm," *J.R. Statist. Soc.*, vol. B39, pp. 1-38, 1977.

[5] B. M. Shahshahani and D. A. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon," *IEEE Trans. on Geosci. Remote Sensing*, vol. 32, no. 5, pp. 1087-1095, September 1994.

[6] J. P. Hoffbeck and David A. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, no. 7, pp. 763-767, July 1996.

[7] J. P. Hoffbeck and David A. Landgrebe, "Classification of Remote Sensing Images having High Spectral Resolution," Remote Sensing of Environment, Vol. 57, No. 3, pp. 119-126, September 1996.

[8] K. Fukunaga and R. R. Hayes, "Effects of sample size in classifier design," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-11, No. 8, pp. 873-885, Aug. 1989.

[9] L. Jimenez and D. A. Landgrebe, "Projection Pursuit in High Dimensional Data Reduction: Initial Conditions, Feature Selection and the Assumption of Normality," IEEE International Conference on Systems, Man, and Cybernetics, Vancouver, Canada, October 22-25, 1995.

Presented at the SPIE International Symposium on Optical Science, Engineering, and Instrumentation, San Diego, CA, July 19-24, 1988