Covariance Estimation With Limited Training Samples

Saldju Tadjudin and David A. Landgrebe
School of Electrical and Computer Engineering
Purdue University West Lafayette, IN 47907-1285
Phone (765) 494-3486 Fax (765) 494-3358 landgreb@ecn.purdue.edu

ABSTRACT

This paper describes a covariance estimator formulated under an empirical Bayesian setting to mitigate the problem of limited training samples in the Gaussian maximum likelihood classification for remote sensing. The most suitable covariance mixture is selected by maximizing the average leave-one-out log likelihood. Experimental results using AVIRIS data are presented.

Index Terms: Gaussian Maximum Likelihood, regularization, covariance estimation.

INTRODUCTION

In the conventional Gaussian maximum likelihood (ML) classifier, the classification rule can be expressed in the form of a discriminant function and a sample is assigned to the class with the largest discriminant function value. A multivariate Gaussian density function is given as

$$f_i(x) = (2\pi)^{-p/2}|\Sigma_i|^{-1/2}\, exp\left[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right] \qquad 1 \le i \le L$$

where $x \in \Re^p$, μ_i and Σ_i are the ith class mean vector and covariance matrix, respectively, and L is the number of classes. Assuming a [0,1] loss function, the maximum likelihood classification rule then becomes

$$d_{\hat{i}}(x) = \min_{1\le i\le L} d_i(x)$$

where d_i is the discriminant function given by

Work leading to this paper was supported in part by NASA under Grant NAG5-3975 and the Army Research Office under Grant DAAH04-96-1-0444.

$$d_i(x) = (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) + ln|\Sigma_i| .$$

This classification rule is also called a quadratic classifier. A special case occurs when all of the class covariance matrices are identical. It then becomes a linear classifier:

$$\Sigma_i = \Sigma \qquad 1 \le i \le L .$$

In practical situations, the true class distributions are rarely known. Therefore, the sample estimates are computed from the training samples.

The quadratic classifier's performance can be degraded when the number of dimensions is large compared to the training set size due to the instability of sample estimates. In particular, the sample covariance estimate becomes highly variable and may even be singular. One way to deal with the instability of covariance estimate is to employ the linear classifier. By replacing each class covariance estimate with their average, leading to the linear classifier, the number of parameters is reduced and thus the variances of their estimates become smaller. Even though each class covariance matrix may differ substantially, studies [1][2] have shown that the decrease in variances of the parameter estimates accomplished by using the linear classifier often leads to better classification performance than the quadratic classifier for small training sample size.

Although a linear classifier often performs better than a quadratic classifier for small training set size, the choice between linear and quadratic classifiers is rather restrictive. Several methods [3][4][5] have been proposed where the sample covariance estimate is replaced by partially pooled covariance matrices of various forms. In this formulation, some degree of regularization is applied to reduce the number of parameters to be estimated, thus improving classification performance with small training set size. Therefore, regularization techniques can also be viewed as choosing an intermediate classifier between the linear and quadratic classifiers.

In general, regularization procedures can be divided into two tasks: 1) the choice of covariance mixture models, and 2) model selection. To perform regularization, one must first decide upon a set of appropriate covariance mixture models that represent a "plausible" set of covariance estimates. Normally, a covariance mixture of the following form is assumed:

$$\hat{\Sigma}_i = (1 - \alpha_i) S_i + \alpha_i S_p \qquad 0 \le \alpha_i \le 1 .$$

The regularization or mixing parameter α_i then controls the biasing of individual class covariance sample estimate S_i to a pooled covariance matrix S_p. However, this partially pooled covariance estimate may not provide enough regularization even for a linear classifier. In the case when the total number of training samples is comparable to or is less than the dimensionality, even the linear classifier becomes ill- or poorly-posed. Therefore, an alternative covariance mixture is provided by biasing the sample covariance toward some non-singular diagonal matrix Λ :

$$\hat{\Sigma}_i = (1-\alpha_i)S_i + \alpha_i\Lambda \qquad 0 \le \alpha_i \le 1$$

For given value(s) of the mixing parameter(s), the amount of bias will depend on how closely the estimates $\hat{\Sigma}_i$ actually represent those true parameters Σ_i. Therefore, the goal of model selection is to select appropriate values for the mixing parameters that can be estimated from minimizing a loss function based on the training samples.

A popular minimization criterion is based on the cross-validated estimation of classification error. Although using this criterion to select the mixing parameters has the benefit of being directly related to classification accuracy, it has some disadvantages as well. First of all, it is computationally intensive. Second, the same mixing parameter has to be used for all classes since the classification procedure requires all covariance estimates simultaneously. However, the same choice of mixing parameter might not be optimal for all classes. Furthermore, the same classification error rate might occur along a wide range of parameter values and hence the optimal value of mixing parameter is non-unique. Therefore, a tie-breaking technique is needed. No studies have indicated the best method for tie-breaking.

Another maximization criterion that has been applied is the sum of the average leave-one-out likelihood values. This criterion requires less computation than the leave-one-out classification error procedure. It also has the advantage that each class covariance matrix can be estimated independently of the others. Therefore, the mixing parameter can be different for each class. Moreover, not all classes need to be subjected to regularization, especially those with sufficient training samples. However, a major drawback of this criterion is the lack of direct relationship with classification accuracy.

PREVIOUS WORK

Friedman [3] has proposed a procedure called "regularized discriminant analysis" (RDA) which is a two-dimensional optimization over covariance mixtures as shown in the following:

$$\hat{\Sigma}_i(\lambda,\gamma) = (1-\gamma)\hat{\Sigma}_i(\lambda) + \gamma\left(\frac{tr\left(\hat{\Sigma}_i(\lambda)\right)}{p}\right)I \qquad 0 \le \gamma \le 1$$

where

$$\hat{\Sigma}_i(\lambda) = \frac{(1-\lambda)(N_i - 1)S_i + \lambda(N-L)S_w}{(1-\lambda)N_i + \lambda N} \qquad 0 \le \lambda \le 1,$$

I is the identity matrix, N is the total number of training samples, and S_w is the average of the sample covariance estimates given as

$$S_w = \frac{(N_1 - 1)S_1 + (N_2 - 1)S_2 + \cdots + (N_L - 1)S_L}{N - L}\ .$$

The regularization parameters are given by the pair (λ,γ), which are obtained by minimizing the leave-one-out cross-validation errors. As mentioned previously, the bias toward a diagonal matrix helps stabilize the covariance estimate even when the linear classifier is ill- or poorly-posed. Furthermore, choosing the diagonal form to be the average eigenvalue times the identity matrix has the effect of decreasing the larger eigenvalues and increasing the smaller ones, thereby counteracting the bias inherent in sample-based estimation of eigenvalues. This diagonal form is also advantageous when the true covariance matrices are some multiples of the identity matrix.

In [4], the covariance matrix is determined from the following pair-wise mixtures: diagonal sample covariance-sample covariance, sample covariance-common covariance, and common covariance-diagonal common covariance matrices. Thus, the estimator has the following form:

$$\hat{\Sigma}_i(\alpha_i) = \begin{cases} (1-\alpha_i)diag(S_i) + \alpha_i S_i & 0 \le \alpha_i \le 1 \\ (2-\alpha_i)S_i + (\alpha_i - 1)S & 1 < \alpha_i \le 2 \\ (3-\alpha_i)S + (\alpha_i - 2)diag(S) & 2 < \alpha_i \le 3 \end{cases}$$

where

$$S = \frac{1}{L} \sum_{i=1}^{L} S_i \, .$$

The variable α_i is the mixing parameter that determines which estimate or mixture of estimates is selected so that the best fit to the training samples is achieved by maximizing the average leave-one-out log likelihood of each class:

$$LOOL_i = \frac{1}{N_i} \sum_{k=1}^{N_i} ln\left[f\left(x_{i,k} \middle| m_{i\backslash k}, \hat{\Sigma}_{i\backslash k}(\alpha_i) \right) \right]$$

where sample k from class i is removed. Once the appropriate value of α_i has been estimated, the estimated covariance matrix is computed with all the training samples and is used in the Gaussian maximum likelihood classifier. Using an approximation on the diagonal matrices, LOOC also requires less computation than RDA. However, without the approximation, LOOC is more computationally expensive than RDA. Also, the average leave-one-out likelihood has no direct relationship to classification accuracy.

An empirical Bayesian method [5] was suggested in which the Σ_i are modeled as the outcomes of a common inverted Wishart prior distribution. The form of covariance mixtures is similar to those in RDA except for the pooled covariance estimate which is formulated under the Bayesian context. The optimal values for (λ, γ) are selected by maximizing the sum of the average leave-one-out class likelihood.

A NEW COVARIANCE ESTIMATOR

In this section, we propose a new covariance estimator based on a Bayesian formulation. The proposed estimator is essentially an extension of previous works in [3][4][5]. The first form of covariance mixtures is derived by assuming that the total number of training samples is greater than the dimensionality. In this case, the common covariance matrix is non-singular.

The assumption of normally distributed samples implies that the sample covariance matrices S_i are mutually independent with Wishart distribution:

$$S_i \sim W\left(\frac{1}{f_i} \Sigma_i, f_i \right)$$

where $f_i = N_i - 1$, N_i is the number of training samples for class i and W denotes the central Wishart distribution with f_i degrees of freedom and parameter matrix Σ_i. Then the family of inverted Wishart distributions provides a convenient family of prior distributions for the true covariance Σ_i. Assume that each Σ_i has an inverted Wishart prior distribution so that the Σ_i are mutually independent with

$$\Sigma_i \sim W^{-1}\left((t-p-1)\Psi,t\right) \qquad t > p+1$$

where W^{-1} is an inverted Wishart distribution with parameters Ψ and t for p dimensions. The prior mean Ψ represents the central location of the prior distribution of the Σ_i, and t controls the concentration of the Σ_i around Ψ.

Under squared error loss, the Bayes estimator of Σ_i is given by [5]

$$\hat{\Sigma}_i(\Psi,t) = \frac{f_i}{f_i+t-p-1}S_i + \frac{t-p-1}{f_i+t-p-1}\Psi .$$

By letting $\alpha_i = \dfrac{t-p-1}{f_i+t-p-1}$ and Ψ be a pooled covariance estimate S_p, the Σ_i can then be replaced by partially pooled estimates of the form:

$$\hat{\Sigma}_i = (1-\alpha_i)S_i + \alpha_i S_p \qquad 0 \le \alpha_i \le 1 .$$

The value of t can in turn be expressed in terms of α_i:

$$t = \frac{\alpha_i(f_i-p-1)+p+1}{1-\alpha_i} \qquad 0 \le \alpha_i < 1 .$$

The pooled covariance estimate is then defined by the generalized least squared estimator of Ψ, designated as $S_p^*(t)$, for L classes and a given t:

$$S_p^*(t) = \left(\sum_{i=1}^{L}\frac{f_i}{f_i+t-p-1}\right)^{-1}\sum_{i=1}^{L}\frac{f_i}{f_i+t-p-1}S_i .$$

When the total number of training samples is close to or less than the number of features, even the pooled covariance matrix becomes unstable. In this case, biasing the sample and common covariance estimates towards some form of diagonal matrix can avoid the problem of singularity. We bias the sample and common covariance estimates towards their own diagonal elements which is advantageous when the class covariance matrix is ellipsoidal. The proposed covariance estimator then has the following form:

$$\hat{\Sigma}_i(\alpha_i) = \begin{cases} (1-\alpha_i)diag(S_i) + \alpha_i S_i & 0 \le \alpha_i < 1 \\ (2-\alpha_i)S_i + (\alpha_i - 1)S_p^*(t) & 1 \le \alpha_i < 2 \\ (3-\alpha_i)S + (\alpha_i - 2)diag(S) & 2 \le \alpha_i \le 3 \end{cases}$$

where $S = \frac{1}{L}\sum_{i=1}^{L} S_i$. The maximization of leave-one-out average log likelihood is used as the criterion to select the appropriate mixture model. Therefore, to select an appropriate mixture, the value of α_i is fixed and the leave-one-out average likelihood is computed and compared for each α_i. The direct implementation of the leave-one-out likelihood function for each class with N_i training samples would require the computation of N_i matrix inverses and determinants at each value of α_i. Fortunately, a more efficient implementation can be derived using the rank-one down-date of the covariance matrix [6].

EXPERIMENTAL RESULTS

For the experiment, we use an AVIRIS data set with 145 X 145 pixels as shown in Fig. 1. The AVIRIS data was taken over NW Indiana's Indian Pine test site in June 1992. The water absorption bands (104-108, 150-163, 220) have been discarded, leaving a total of 200 channels. This data contains 17 classes of varying sizes. The ground truth map is shown in Fig. 2. The purpose of this experiment is to demonstrate the effect of covariance estimation on classes with varying covariance structures and different training sample sizes. The training samples are selected to be 20% of the number of labeled samples for each class. The labeled samples, excluding the training samples are then used as test samples. The classes and the numbers of labeled samples are listed in Table 1.

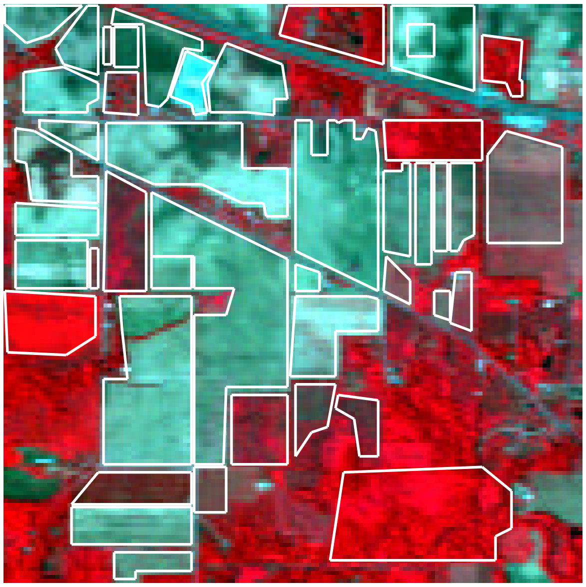

Fig. 1. AVIRIS Data Scene (Original in Color. A color version of all figures may be viewed at http://dynamo.ecn.purdue.edu/~landgreb/publications.html).

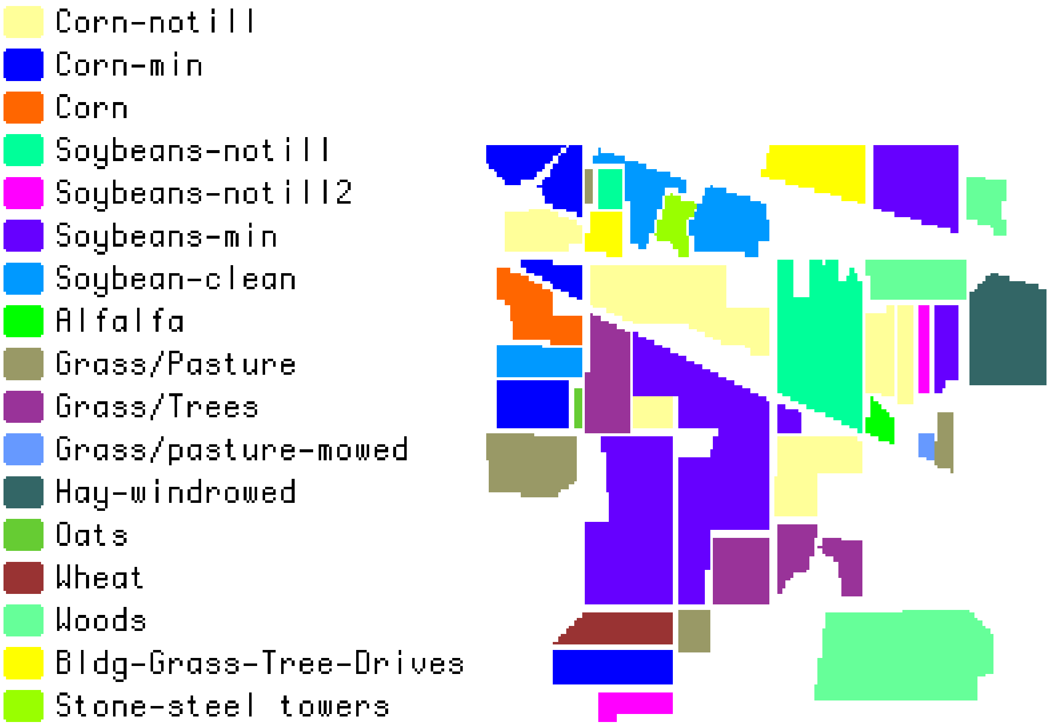

Fig. 2. AVIRIS Data Ground Truth Map (Original in Color)

Class Name	No. of Labeled Samples
1. Corn-no till	1423
2. Corn-min till	834
3. Corn	234
4. Soybeans-no till	797
5. Soybeans-no till2	171
6. Soybeans-min till	2468
7. Soybeans-clean till	614
8. Alfalfa	54
9. Grass/Pasture	497
10. Grass/Trees	747
11. Grass/pasture-mowed	26
12. Hay-windrowed	489
13. Oats	20
14. Wheat	212
15. Woods	1294
16. Bldg-Grass-Tree-Drives	380
17. Stone-steel towers	95

Table 1. Class Description of the AVIRIS Data.

This data was obtained in June 1992 so most of the row crops in the agricultural portion of the test site had not reached their maximum ground cover. Therefore, the classification of these crops becomes challenging since the spectral information comes from a mixture of the crops, the soil variations and previous crop residues. These crops are listed as the first seven classes and their mean classification accuracy is computed separately.

The classification procedures for testing the data are shown in Table 2. Since the Euclidean distance classifier does not utilize the covariance information, its performance would indicate whether the second order statistics are useful for the classification of high dimensional data with limited training samples. The use of the common covariance estimate for all classes is equivalent to a linear classifier. The leave-one-out covariance estimator (LOOC) [4] is implemented to compare with the proposed Bayesian leave-one-out covariance estimator (bLOOC). The mixing parameter α_i is set at 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75 and 3. In addition to using the covariance estimator to help increase the stability of covariance estimate, feature extraction can also help reduce the number of features to cope with small training set sizes. We perform Discriminant Analysis Feature Extraction [8] (DAFE) which only utilizes mean information and is therefore less sensitive to small training sample size. The sample covariance estimate is not tested in this experiment since the numbers of training samples for some classes are extremely small. Two types of classifiers, namely, the quadratic classifier (QC) and the spatial-spectral classifier ECHO [9] (Extraction and Classification of Homogeneous Objects) are then applied and compared. While the quadratic classifier assigns individual pixels to one of the classes, the ECHO classifier first divides the image into groups of contiguous pixels and classifies each group to one of the classes. The results of classification are shown in Table 2 and Fig. 3. The highest accuracy is highlighted in bold letters.

Procedures	Classes 1-17 (%)	Classes 1-7 (%)
1. Euclidean Distance	48.2	31.8
2. Common Cov+DAFE+QC	74.8	70.2
3. Common Cov+DAFE+ECHO	76.8	74.9
4. LOOC+DAFE+QC	75.3	70.7
5. LOOC+DAFE+ECHO	80.4	82.7
6. bLOOC+DAFE+QC	75.5	72.6
7. bLOOC+DAFE+ECHO	**82.9**	**89.1**

Table 2. Classification Accuracy for the AVIRIS Data.

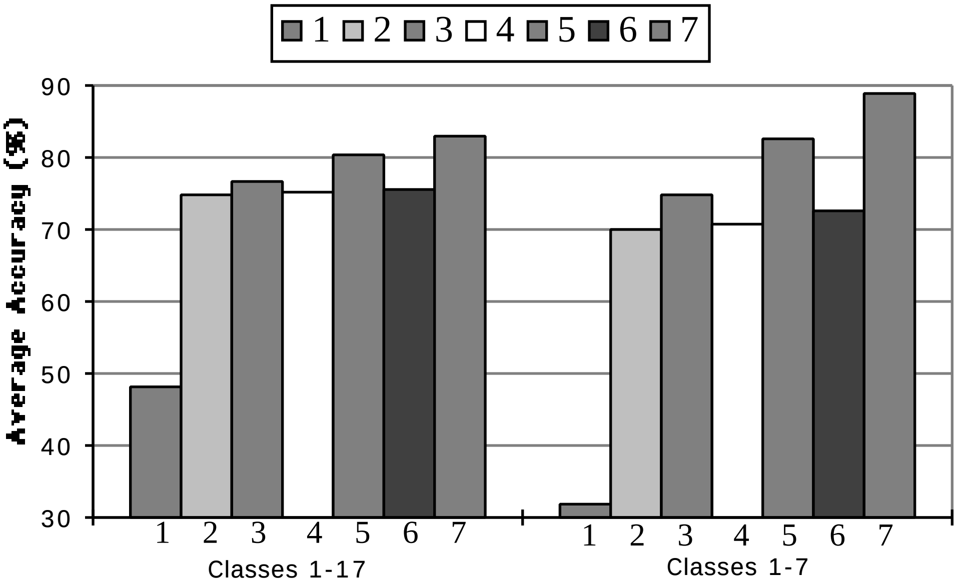

Fig. 3. Mean Classification Accuracy

DISCUSSION

The performance of the Euclidean distance classifier is significantly lower than the other classifiers. This shows that the second order statistics are useful for classifying these high dimensional data even though the training samples are limited. Although the class covariance matrices differ substantially, the use of common covariance matrix and hence the linear classifier improves the performance substantially compared to the Euclidean distance classifier. The table shows that the best performance is achieved by using bLOOC, DAFE and the ECHO classifier. The classification accuracy increases substantially for the row crops 1-7. Compared with the second best result obtained from the classifier LOOC+DAFE+ECHO, the accuracy increases from 82.7% to 89.1%. The mean accuracy for all classes improves from 80.4% to 82.9% as well. It should be mentioned that when all classes have equal number of training samples, bLOOC has the same form as LOOC. Therefore, the proposed Bayesian estimator is beneficial when the sample sizes are unequal and the training set size reflects the true priors. The classification maps for LOOC+DAFE+ECHO and bLOOC+DAFE+ECHO are shown in Fig. 4 and 5, respectively.

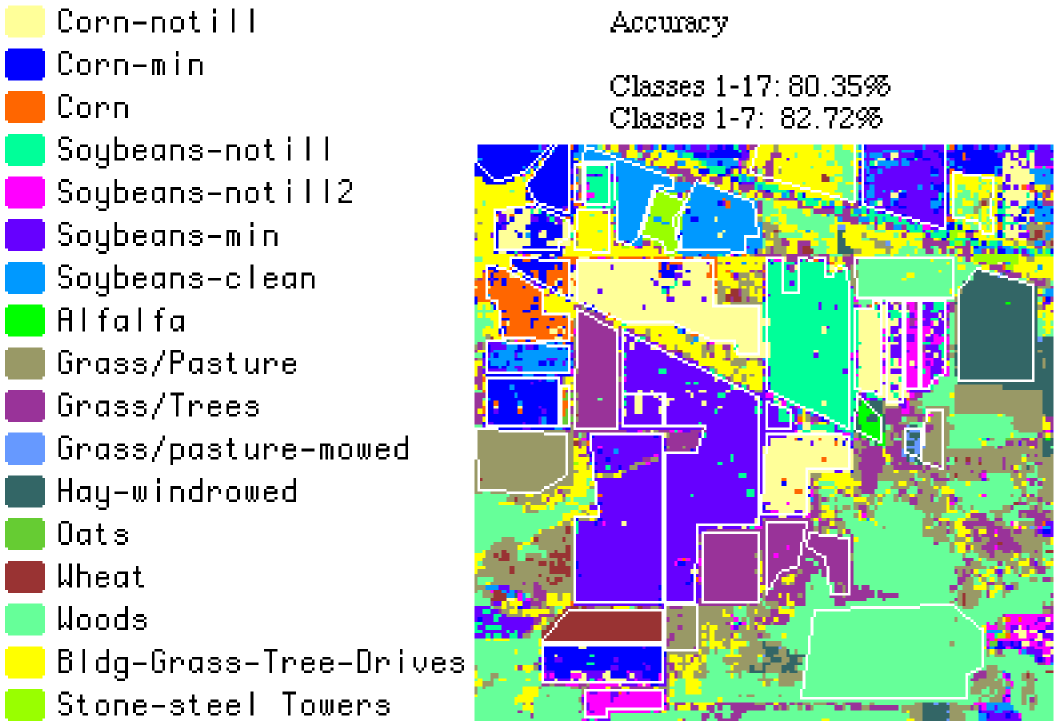

Fig. 4. Classification Map for Method: LOOC+DAFE+ECHO (Original in Color)

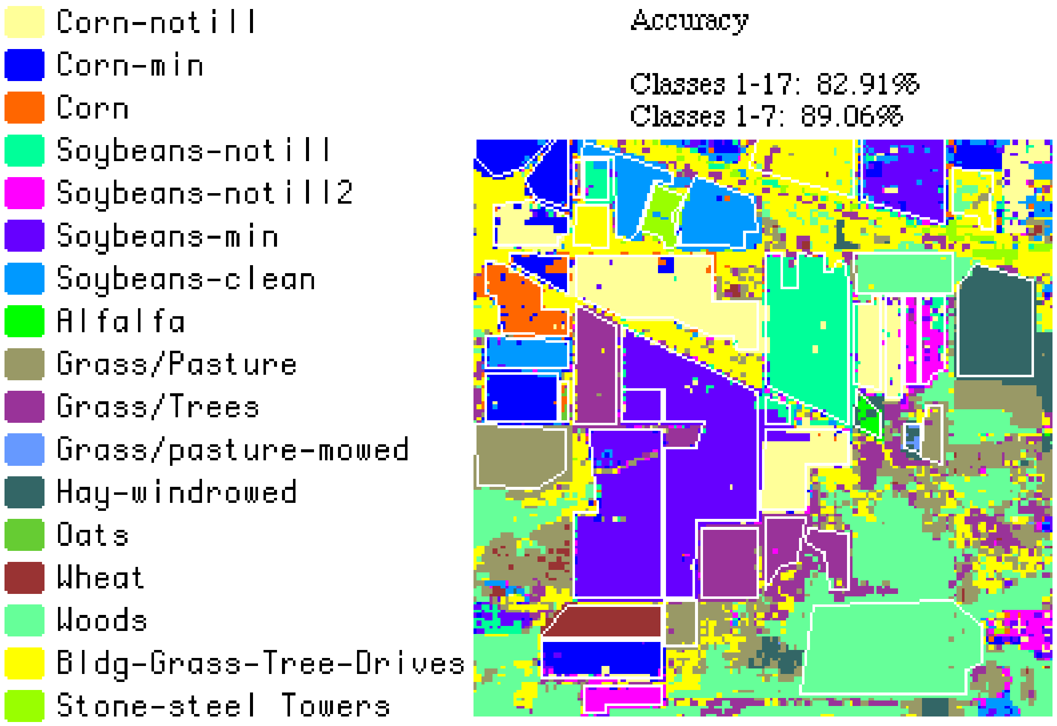

Fig. 5. Classification Map for Method: bLOOC+DAFE+ECHO (Original in Color)

CONCLUSION

The inverse of a covariance matrix becomes ill- or poorly-posed if the training set size is small compared to the dimensionality. Conventionally, the stabilization of the covariance estimate has been accomplished by regularization, which tends to reduce the variance of the estimate at the expense of increased bias. This method can also be viewed as a compromise between the linear and quadratic classifiers. In this paper, a regularization method under the Bayesian setting has been proposed. The proposed Bayesian leave-one-out covariance (bLOOC) estimation method was shown to have better performance than other methods when the training set size reflects the true priors of the classes. This is particularly true for remote sensing applications since more training samples are usually selected for larger classes. When used in conjunction with discriminant analysis feature extraction (DAFE), the proposed covariance estimation was demonstrated to circumvent the limited training set size problem. However, since the leave-one-out likelihood is used as the criterion for the estimator, it has the drawback of not being directly related to class separability, and subsequently the classification accuracy. Therefore, some smooth loss function derived from the class separability is recommended for future work. Also, since DAFE does not work well when the classes have similar mean values, alternative feature extraction or classification methods need to be explored.

REFERENCES

[1] S.P. Lin and M.D. Perlman, "A Monte Carlo comparison of four estimators of a covariance matrix," Multivariate analysis--VI : Proceedings of the Sixth International Symposium on Multivariate Analysis, P.R. Krishnaiah, ed., Amsterdam: Elsevier Science Pub. Co., 1985, pp. 411-429.

[2] P.W. Wahl and R.A. Kronmall, “Discriminant functions when covariances are equal and sample sizes are moderate,” Biometrics, Vol. 33, pp. 479-484, 1977.

[3] J.F. Friedman, "Regularized discriminant analysis," J. R. Statist. Soc., Vol 84, pp. 17-42, 1989.

[4] J.P. Hoffbeck and D.A. Landgrebe, "Covariance matrix estimation and classification with limited training data", IEEE Transaction of Pattern Analysis and Machine Intelligence, Vol 18, No. 7, pp. 763-767, 1996.

[5] W. Rayens and T. Greene, "Covariance pooling and stabilization for classification," Computational Statistics and Data Analysis, Vol 11 pp. 17-42, 1991.

[6] S. Tadjudin, Classification of High Dimensional Data with Limited Training Samples, Ph.D. Thesis, School of Electrical and Computer Engineering, Purdue University, 1998 (123 pages). Available from http://dynamo.ecn.purdue.edu/~landgreb/publications.html

[7] C. Lee and D.A. Landgrebe, “Feature extraction based on decision boundaries,” IEEE Transaction of Pattern Analysis and Machine Intelligence, Vol. 15, No. 3, pp. 388-400, 1993.

[8] K. Fukunaga, Introduction to Statistical Pattern Recognition. 2nd Ed., Boston: Academic Press, 1990.

[9] R.L. Kettig and D.A. Landgrebe, "Classification of multispectral image data by extraction and classification of homogeneous objects," IEEE Trans. Geosci. Electro., Vol. GE-14, No. 1, pp. 19-26, 1976.