RANDOM FORESTS OF BINARY HIERARCHICAL CLASSIFIERS FOR ANALYSIS OF HYPERSPECTRAL DATA

Melba M. Crawford[1*], JiSoo Ham[1], Yangchi Chen[1], Joydeep Ghosh[2]
[1]Center for Space Research, 3925 W. Braker Lane, Austin, TX 78759, [*]email: crawford@csr.utexas.edu
[2]Department of Electrical and Computer Engineering
The University of Texas at Austin

Abstract – **Statistical classification of hyperspectral data is challenging because the input space is high in dimension and correlated, but labeled information to characterize the class distributions is typically sparse. The resulting classifiers are often unstable and have poor generalization. A new approach that is based on the concept of random forests of classifiers and implemented within a multiclassifier system arranged as a binary hierarchy is proposed. The primary goal is to achieve improved generalization of the classifier in analysis of hyperspectral data, particularly when the quantity of training data is limited. The new classifier incorporates bagging of training samples and adaptive random subspace feature selection with the Binary Hierarchical Classifier (BHC), such that the number of features that is selected at each node of the tree is dependent on the quantity of associated training data. Classification results from experiments on data acquired by the Hyperion sensor on the NASA EO-1 satellite over the Okavango Delta of Botswana are superior to those from our original best basis BHC algorithm, a random subspace extension of the BHC, and a random forest implementation using the CART classifier.**

Keywords – **binary hierarchical classifier, classification, EO-1, Hyperion, hyperspectral, Okavango Delta, random forests, random subspace feature selection.**

1. INTRODUCTION

The increasing availability of data from hyperspectral sensors, particularly with the launch of the Hyperion instrument on the NASA EO-1 satellite, has generated tremendous interest in the remote sensing community. These instruments characterize spectral signatures with much greater detail than traditional multispectral sensors and thereby can potentially provide improved discrimination of targets [1]. However, hyperspectral data also present difficult challenges for supervised statistical classification, where labeled training data are used to estimate the parameters of the label-conditional probability density functions [2]. The dimensionality of the data is high (~200), there are often tens of classes *C*, and the quantity of training data is often small. Sample statistics of training data may also not be representative of the true probability distributions of the individual class signatures, particularly for remote, inaccessible areas where training data are logistically difficult and expensive to acquire. Generalization of the resulting classifiers is often poor, thereby resulting in poor quality mapping over extended areas. Various approaches have been investigated to mitigate the impact of these three issues.

Small sample problems. The substantial methodology in this area can be largely categorized as one of three approaches [3]. Regularization methods, including "shrinkage," try to stabilize the estimated covariance matrix directly by weighting the sample covariance matrix as well as "supplemental" matrices [4]. The covariance matrix can be "shrunk" toward the identity matrix or a pooled covariance matrix. Hybrid approaches assign weights to the sample covariance (normal and diagonal) matrix and a pooled covariance matrix [5]. While this may reduce the variance of the parameter estimates, the bias of the estimates can increase dramatically.

Alternatively, the input space can be transformed into a reduced feature space via feature selection, feature extraction, or artificially adding labeled samples. Feature subset selection methods [6-9] may provide valuable domain knowledge about the importance of inputs, but are often implemented as sub-optimal greedy algorithms. They are also sensitive to anomalies in the training data, particularly for small training samples, and thus may not yield robust classifiers with good generalization. They also do not exploit the redundancy exhibited in hyperspectral data. Specific techniques for identifying and augmenting the existing training data with unlabeled data have also been developed and shown to enhance supervised classification [10-13]. However, convergence of the updating scheme can be problematic, and it is affected by selection of the initial training samples and by outliers.

Extraction methods such as the principal component transformation or the Fisher discriminant may result in some loss of interpretability and can be poorly estimated due to limited training data. In analysis of hyperspectral data, Lee and Landgrebe proposed methods for feature extraction from hyperspectral data based on decision boundaries that

maximize separation of data in multiple two-class problems [14]. These decision boundary feature extraction (DBFE) methods are often effective for two-class problems, but do not exploit correlation between sequential bands. Jia and Richards developed the Segmented Principal Components Transformation (SPCT) whereby the original bands are grouped into subsets of highly correlated adjacent bands to which the K-L transform is applied. The most significant principal components are then selected from each subset to yield a feature vector with reduced dimension [15]. The approach treats inter-band correlation globally and does not guarantee good discrimination capability because the PCT preserves variance in the data rather than maximizing discrimination between classes. Kumar et al. investigated band combining techniques, motivated by best-basis functions, as a means of feature extraction in a pairwise classifier framework [16]. Adjacent bands are selected for merging (alt. splitting) in a bottom-up (alt. top down) fashion using the product of a correlation measure and a Fisher discriminant. Morgan et al. [17] suggested a similar correlation-based band combining approach, in conjunction with a covariance shrinkage method, for both a top-down and bottom up hierarchical classifier to ameliorate the small training data problem.

The third approach uses an ensemble of "weaker" classifiers. Bagging involves bootstrapped sampling of the original data and generating a classifier specific to each sample [18]. When the training data set in the (sub -)sample is very small, the potential for improved diversity and reduced impact of outliers is offset by degradation in individual classifier performance [19]. Boosting also combines weak individual classifiers to develop an improved classifier, but by re-weighting training data to increase sensitivity to incorrectly classified training observations. While boosting can improve performance for large training samples, it not useful for small sample problems, particularly in the presence of outliers. When the input space is large, random subspace (RS) feature selection can potentially provide improved classifier diversity, while stabilizing parameter estimates, by randomly reducing the number of inputs to each classifier in the ensemble and constructing multiple classifiers in the resulting random input space [20], [21]. The method is potentially attractive for problems with redundant input features (e.g. hyperspectral data) and when outliers exist in the training data. Recently, approaches referred to as "random forests of classifiers" involve developing forests trees from randomly sampled subspaces of input features, then combining the resulting independent outputs via voting or a maximum *a posteriori* rule [22]. These methods typically achieve superior generalization for small training samples, but are computationally intensive and interpretability may be limited.

Large output space problems. Output decomposition using binary classifiers in a multiclassifier framework has been shown to be more successful than traditional 1-of-*C* classifiers for many problems [23]. Decomposition methods using pairwise classifiers [16][24][25], error correcting output codes (ECOC) [26], and binary decision trees [27][28] have all been investigated in this context. Pairwise classifiers develop a separate classifier for each pair of classes, thereby resulting in $O\left(C^2\right)$ classifiers which must be combined to determine the final class label. These methods often yield simple classifiers with excellent discrimination for specific pairs, but are generally inefficient for problems with a large number of output classes. They also do not exploit natural groupings of classes, which can improve classifier robustness and transferability.

In the ECOC, a *C*-class problem is decomposed into $\overline{C}$ binary problems whereby the original class is then encoded into a $\overline{C}$ binary vector of a coding matrix. Novel observations are labeled as members of the class whose codeword is closest to that formed by the outputs of the $\overline{C}$ classifiers. Similar to the pairwise classifiers, the binary structure mitigates the impact of small training samples and often yields robust, stable classifiers, but the ECOC may be problematic for large numbers of output classes as the length of a code associated with the binary classifiers $\overline{C}$ can be quite large. Further, the code matrix design is not based on the characteristics of the classes it represents, thereby limiting interpretability of the classifier.

Binary trees, which often provide an attractive approach for decomposing large output space problems, can be constructed using a variety of splitting functions involving single or multiple features and output classes. To address the high dimensional output problem while exploiting the affinity for spectrally similar classes, Kumar et al. proposed a Binary Hierarchical Classifier (BHC) [29] to decompose a $(C > 2)$-class problem into a binary hierarchy of $(C-1)$ simpler 2-class problems that can be solved using a corresponding hierarchy of classifiers, each based on a simple linear discriminant. The method was extended by Morgan et al. [17] for small training samples using an adaptive best basis BHC, which exploits the class specific correlation structure between sequential bands of hyperspectral data and utilizes an adaptive regularization approach to stabilize covariance estimates. An adaptive random subspace feature selection approach was also investigated within the BHC framework (RS BHC) as a means of improving classifier performance when the number of training samples is extremely small [30].

In this paper, we investigate a random forest of binary classifiers as a means of further increasing diversity of the hierarchical classifiers produced by the BHC. Our goal is to use the BHC structure to exploit the advantages of natural class affinity while improving generalization in classification of hyperspectral data when the number of training samples is small. A secondary goal is to mitigate the impact of sensor noise and residual atmospheric artifacts on the classification results. The paper is organized as follows: the BHC method, including the best basis (BB BHC), random subspace (RS BHC), and random forest

implementations (RF BHC), is described in Section II; classification results for test and independent test sets of data acquired by the NASA EO-1 Hyperion sensor for mapping land cover in the Okavango Delta of Botswana are presented in Section III and compared to those obtained from the BB BHC, RS BHC, and a random forest implementation using CART [22] [27]; results from all the methods are evaluated, and new directions for future work are discussed in Section IV.

II. RANDOM FOREST BINARY HIERARCHICAL CLASSIFICATION METHOD

The top-down Binary Hierarchical Classifier (BHC) framework recursively decomposes a *C*-class problem into *C*-1 two-(meta)class problems via a deterministic simulated annealing method [29]. The root classifier tries to optimally partition the original set of classes into two disjoint meta-classes while simultaneously determining the Fisher discriminant that separates these two subsets. This procedure is recursed, i.e., the meta-class Ω_n at node *n* is partitioned into two meta-classes $(\Omega_{2n}, \Omega_{2n+1})$, until the original *C* classes are obtained at the leaves. The tree structure, as shown in Figure 1, allows the more natural, easier discriminations to be accomplished earlier.

While the BHC exploits natural class affinities, it is affected at lower levels of the hierarchy if the input space is large and the number of training samples is small. The BB-BHC ameliorates this effect by utilizing an ancestor covariance matrix while exploiting the inter-band serial correlation through an adaptive, class dependent, band aggregation process [17]. A band combining step is performed on highly correlated, spectrally adjacent bands prior to the partitioning of meta-classes, thereby reducing the number of inputs relative to the number of training data points. Bands are aggregated until a user defined ratio, *R*, between the number of training data for the respective (meta)classes and input dimension is achieved. Typically, *R* is selected to be at least 5.

The RS BHC method extends the BHC by utilizing the random subspace method as a second phase to reduce the actual number of inputs while sharpening classifier boundaries [31]. The BB BHC method is used to first construct the hierarchy, then random subspace sampling is performed at each node of the tree where the ratio *R* is not satisfied. For each (meta) class *m* with n_m vector-valued observations, $\boldsymbol{X}_m = (\boldsymbol{X}_1, ..., \boldsymbol{X}_{n_m})$, a subset of elements of $\boldsymbol{X}_i = (x_{i1}, ...x_{ik})$ with dimension $p_m = n_m / R < k$ is then randomly selected from the *k*-dimensional set of features. The resulting modified training set $\boldsymbol{X}_m^r = (\boldsymbol{X}_1^r, ..., \boldsymbol{X}_{n_m}^r)$ consists of observation vectors, $\boldsymbol{X}_i^r = (x_{i1}^r, ...x_{ip}^r)$, where the same subset of features is selected for each element $\boldsymbol{X}_i^r \in \boldsymbol{X}^r, (i = 1, ...n_m)$. Classifiers are constructed for each random subspace, and results are typically combined at each node of the hierarchy via majority voting. The number of random subspaces selected at each such node is $N_s = (k / p_m) \times F$, where value of F is a user supplied input. Our empirical evidence indicates good results are achieved for $2 < F < 4$, but improvement in classification accuracy is not significant for $F > 4$. A best basis version (BB RS BHC) of the method also incorporates band aggregation in the random subspace phase, but terminates the band combining when correlation between successive band groups falls below a specified threshold.

The random forest implementation of the BHC (RF BHC) incorporates random subspace feature selection in the actual development of the tree, whereas the RS BHC method uses it only as a means to reduce the input space and refine the decision boundaries obtained by the BB BHC. For the BHC, this is particularly advantageous as random subspace sampling is performed only at nodes where the ratio, *R,* is not exceeded. Thus, sub-sampling of the input features typically occurs only at lower levels of the tree. For moderate sized training samples, bagging can thus increase diversity of the multiclassifier system. For each tree in the RF BHC, a bootstrap sample of observations is selected. At each metaclass node *m*, a random subspace of features of dimension $p_m = n_m / R$, is selected if $p_m < k$. Otherwise, the full feature set is used to determine the decision boundary for the classifier at that node. This guarantees that the number of input features selected at each node automatically satisfies the ratio criterion, *R*. The tree is then developed using the resulting set of features selected at each node. The process is repeated to grow a forest of identically, independently distributed random vectors associated with the individual trees. Because the sample sizes at the higher levels of the tree are large, random sub-sampling of features is still deferred to lower levels of the tree. A second version of the method achieves greater diversity by forcing sub-sampling of features at each node of the hierarchy using $p_m = \min(p_m, N_f)$, where N_f is a user selected input. The two implementations are referred to hereafter as the RF BHC1 and RF BHC2 methods, respectively.

III. RESULTS

The NASA EO-1 satellite acquired a sequence of data over the Okavango Delta, Botswana in 2001-2003. The Hyperion sensor on EO-1 acquires data at 30 m^2 pixel resolution over a 7.7 km strip in 242 bands covering the 400-2500 nm portion of the spectrum. Preprocessing of the data was performed by the UT Center for Space Research to mitigate the effects of bad detectors, inter-detector miscalibration, and intermittent anomalies. Uncalibrated and noisy bands that cover water absorption features were removed, and the remaining 145 bands were included as candidate features: [10-55, 82-97, 102-119, 134-164, 187-220].

The data analyzed in this study, acquired May 31, 2001, consist of observations from 14 identified classes

representing the land cover types in seasonal swamps, occasional swamps, and drier woodlands located in the distal portion of the Delta [31]. The class names and corresponding numbers of ground truth observations used in the experiments are listed in Table 1, and the image in Figure 2a shows the complex spatial distribution of classes over the study area. Ten randomly sampled partitions of the training data were sub-sampled such that 75% of the original data were used for training and 25% for testing. In order to investigate the impact of the quantity of training data on classifier performance, these training data were then sub-sampled to obtain ten samples comprised of 50%, 30%, and 15%, of the original training data. All classifiers were evaluated using the ten test samples composed of 25% of the original training data. Because the training and test data are spatially collocated, an independent test set from a nearby region was also acquired and used to evaluate the classifiers. Hereafter, these data are referred to as the test and independent test data, respectively.

Table 1: Class codes, names, and number of training samples for Hyperion data, May 31, 2001

	Class	No. samples
1	water	270
2	hippo grass	101
3	floodplain grasses1	251
4	floodplain grasses2	215
5	reeds1	269
6	riparian	269
7	firescar2	259
8	island interior	203
9	acacia woodlands	314
10	acacia shrublands	248
11	acacia grasslands	305
12	short mopane	181
13	mixed mopane	268
14	exposed soils	95

Experiments were performed using the BB BHC, the random subspace method using both the original inputs and best basis features (RS BHC and BB RS BHC, and the two variations of the random forest BHC method (e.g. RF BHC1 and RF BHC2). Results from a random forest implementation using bagging in conjunction with the CART classifier were included for comparison. The RF CART approach selects a random subspace of features at each node of the tree. The most discriminating feature of the subset is then selected to perform the split.

For the random subspace method, the ratio R was set at 5, and the value of F was set to 4. (Sensitivity analysis indicated that larger values of F do not improve results for this data set.) Although authors recommend various values for the dimension of the random subspace and the number of trees in a random forest, there do not appear to have been any systematic studies of the issue at this time. In our experiments, the dimension of the random subspace was determined adaptively in the BHC, but was always selected such that the ratio of training data to input features, R, was at least 5. For the RF BHC2, the value of N_f was selected to be 20. In order to have somewhat comparable inputs, 20 input features were randomly selected in the RF CART method. 100 trees were grown for each experiment as sensitivity studies showed that larger forests do not provide improved results for this data set.

Figure 2b contains a representative classification result. Average classification accuracies and associated standard deviations for the 10 experiments conducted with each classifier are shown in Figure 3. The overall trends in accuracies relative to the quantity of training data are similar for all methods when applied to the test data set. At the 75% sampling rate, the accuracies for the BHC methods are all quite similar, although the RF BHC1 is somewhat higher and the RF BHC2 somewhat lower than the BB BHC and the RS methods. The results obtained using the BB BHC method consistently have the lowest overall average accuracies. The RS methods yielded similar accuracies to the BB BHC approach at 75% and 50% sampling rates, but improved relative to the BB BHC as sampling rates were reduced. This appears to demonstrate the value of reduced redundancy in the input space and improvements achieved by better tuning of the decision boundaries, even though the tree structure is the same as for the BB BHC and random sampling of the feature space is not required until lower levels of the tree (particularly for the higher training data fractions). Results obtained from using the original and best basis aggregated data with the RS BHC are not statistically different, although the accuracies from using the original data are always somewhat higher. Because improvement was marginal in the RS BHC and the increase in associated computation is substantial, best basis band aggregation was not investigated in conjunction with the random forest methods. The overall accuracies of the random forest methods improve relative to the other BHC methods as the fraction of training data is reduced. The RF BHC1 method is consistently the highest of all methods for the test set, although the difference in average accuracies of the two BHC implementations of the random forest method decreases with smaller training samples. Thus, the test data appear to indicate that there is no increased value in forcing random sampling of the feature space, and thereby inducing greater instability within the forest, at every node of the tree. The standard deviations of the accuracies from all the BHC methods are essentially the same at the 75% sampling rate, but as the sampling rate of training data is reduced, the standard deviations from the BB BHC and RS BHC methods increase approximately linearly, while those from the RF BHC methods remain small and are nearly constant. The overall accuracy of the RF CART method for the test set is lower than that of all the BHC methods by more than two standard deviations at the 75% and 50% sampling rates, and by more than 1 standard deviation at the 30% and 15% sampling rates. This is indicative of the value of the inherent exploitation of class affinities by the BHC approaches. Similar to the RF BHC methods, the standard deviations of the accuracies are approximately constant for

the RF CART approach, although they are higher than for the RF BHC experiments.

Overall accuracies obtained from analysis of the independent test set also support the use random sampling of the input space, but follow different trends. As with the test data, the BB BHC yielded the lowest overall average accuracy at all sampling rates. The incremental improvement in average accuracy achieved by the random subspace method increases with reduced sampling rates, but is not statistically significant as the standard deviations of the accuracies also increase substantially with lower sampling rates. Both of the random forest BHC implementations and the RF CART method yielded higher accuracies for the independent test data at all sampling rates than the BB BHC and RS BHC methods, demonstrating the greater generalization of these approaches. Unlike results from the test data, the RF BHC2 method consistently produced the highest overall average accuracies for the independent test set, indicating the value of the increased diversity of trees achieved by forcing random sampling of the input space at all nodes. Results from the RF CART method degraded with sampling rates below 30%. The overall accuracies of the RF BHC1 actually increase for the independent test set with decreased training sample size, indicating that the reduced quantity of training data is offset by the greater diversity provided by smaller random subsets. This is likely related to both the redundancy of the highly correlated feature space in hyperspectral data and the fact that smaller input spaces should also contain fewer irrelevant features. The difference in overall classification accuracies for the RF BHC1 and RF BHC2 methods is reduced with smaller sampling rates as the two methods converge to a single sampling scheme as forced sampling is eventually not required, even at the top node of the tree.

The class specific accuracies are shown in Figure 4 for the random forest methods. While the performance of the RF BHC2 is generally better than the RF BHC1 and RF CART methods at both 75% and 15% sampling rates, the overall higher classification accuracy at 75% is strongly influenced by its performance for Class 2 and Class 11. Class 2, hippo grass, has a small training sample and its spectral signature in $30m^2$ pixels is quite similar to water as many pixels are mixed with water. Class 11, acacia grasslands, is a mixed class that is most often confused with other grasses or acacia shrubs. At the 15% sampling rate, both BHC methods produce similar results for all classes, while the RF CART method has lower average accuracies with substantially higher standard deviations than the RF BHC methods for most classes. In fact, the higher apparent accuracy of the RF CART method for Class 2 is not statistically significant as the standard deviation of the average sample accuracy is more than 12. While the number of training samples for exposed soils is small, its signature is easily discriminated from that of vegetation, so classification accuracies are consistently high, even at low sampling rates.

IV. CONCLUSIONS AND FUTURE WORK

The primary purpose of the study was to investigate the performance of random feature subset selection methods in terms of generalization. The secondary goal was to investigate the performance of the methods when applied to noisy data sets. An implementation that focused on tuning decision boundaries of the BHC and three random forest approaches were investigated. For the data analyzed in these experiments, the value of the RS BHC extension is marginal in terms of improved classification accuracy of the independent test set. The computational requirements increase substantially, particularly with smaller training sample fractions which require more subset feature selection. Also, the improved SNR from band aggregation is offset by the improved diversity achieved by random sampling of the original features.

The random forest methods all yield superior results for both test and independent test data, with the improvement being greater for the independent test set, thereby indicating improved generalization. The value of the BHC for classes with natural affinities is also demonstrated in the RF BHC implementations relative to the RF CART method. For high sampling rates, the value of increased diversity associated with the RF BHC2 method produced higher classification accuracies. However, as the size of the training sample was reduced, the improved diversity of the RF BHC1 resulting from the smaller feature subsets dominated the forced diversity of the RF BHC2. Additional study is required to better characterize this issue. In this context, elimination of irrelevant and possibly redundant input features should also be considered in the RF BHC. Previous results by Oza and Tumer [32] indicate that this may be a promising direction. Other classifiers should also be investigated within the RF BHC framework. Overall, the RF BHC methods appear to be quite promising in terms of generalization, but should be applied to many more data sets with different characteristics in order to better assess their overall performance.

REFERENCES

[1] J. S. Pearlman, P. S. Berry, C. Segal, J. Shapanski, D. Beiso, and S. Carman, "Hyperion: a space-based spectrometer, *IEEE Trans. Geosci. Rem. Sens.*, in press.

[2] D. Landgrebe, "Hyperspectral image data analysis as a high dimensional signal processing problem," (Invited), Special Issue of the *IEEE Signal Processing Magazine*, **19**(1): 17-28, 2002.

[3] S.J. Raudys and A.K. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners", *IEEE Trans. Pattern Anal. Mach. Intell.*, **13**(3): 252–64, 1991.

[4] M. Skurichina and R.P.W. Duin, "Stabilizing classifiers for very small sample sizes", *Proc. 13th Int. Conf. on Pattern Recognition* (Vienna, Austria, Aug.25-29) vol. 2, Track B: Pattern Recognition and Signal Analysis, IEEE Computer Society Press, Los Alamitos, 891-6, 1996.

[5] S. Tadjudin and D.A. Landgrebe, "Covariance estimation with limited training samples," *IEEE Trans. Geosci. Rem. Sens.*, **37**(4): 2113-8, 1999.

[6] A. Jain and D. Zongker, "Feature selection: evaluation, application and small sample performance," *IEEE Trans. Pattern Anal. Machine Intell.*, **19**(2): 153-158, 1997.

[7] B.S. Serpico and L. Bruzzone, "A new search algorithm for feature selection in hyperspectral remote sensing images," *IEEE Trans. Geosci. Rem. Sens,* **39**(7): 1360-1367, 2001.

[8] A. Henneguelle, *Feature Extraction for Hyperspectral Data Analysis*, Masters Thesis, University of Texas at Austin, 2002.

[9] D. Korycinski, *Investigating the Use of Tabu Search to find Near-Optimal Solutions in Multiclassifier Systems*, PhD. Dissertation, The University of Texas at Austin, 2003.

[10] Q. Jackson and David Landgrebe, "An adaptive classifier design for high-dimensional data analysis with a limited training data set", *IEEE Trans. Geosci.Rem. Sens*, **39**(12): 2664-79, 2001.

[11] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," *Proc. 11th Annual Conf. Computational Learning Theory*, 92-100, 1998.

[12] B. Jeon and D. Landgrebe, "Partially supervised classification using weighted unsupervised clustering," *IEEE Trans. Geosci.Rem. Sens.*, **37**(2): 1073-9, 1999.

[13] 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. Geosci.Rem. Sens.*, **32**(5): 1087-95, 1994.

[14] C. Lee and D. Landgrebe, "Decision boundary feature extraction for neural networks," IEEE Trans. Neural Networks, 8(1): 75-83, 1997.

[15] X. Jia and J.A. Richards, "Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification", *IEEE Trans. Geosci.Rem. Sens.*, **37**(1): 538-42, 1999.

[16] S. Kumar, J. Ghosh, and M.M. Crawford, "Best basis feature extraction algorithms for classification of hyperspectral data," *IEEE Trans. Geosci. Rem. Sens.* **29**(7): 1368-79, 2001.

[17] J.T. Morgan, A. Henneguelle, M.M. Crawford, J. Ghosh, and A. Neuenschwander, "Adaptive feature spaces for land cover classification with limited ground truth," in *Proc. Third Intl. Workshop, MCS 2002,* F. Roli and J. Kittler, Eds. Germany: Springer-Verlag Lecture Notes in Computer Science (#2364), 189-200, 2002.

[18] L. Breiman, "Bagging predictors," *Machine Learning*, **24**(2): 123-40, 1996.

[19] K. Tumer and J. Ghosh, "Error correlation and error reduction in ensemble classifiers" *Connection Science*, Special Issue on Combining, **8**(3/4), 385-404, 1996.

[20] T.K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.,* **20**(8): 832-844, 1998.

[21] M. Skurichina and R.P. W. Duin, "Bagging, boosting, and the random subspace method for linear classifiers," *Int. J. Pattern Analysis and Applications*, **5**(2): 121-135, 2002.

[22] L. Breiman, "Random forests," *Machine Learning*, **45**: 5-32, 2001.

[23] J. Furnkranz, "Round robin classification," *J. Machine Learning Research*, **2**: 721-747, 2002.

[24] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," *Proc. Int. Joint Conf. on Neural Networks*, Washington, D.C., 1999.

[25] M.M. Crawford, S. Kumar, M.R. Ricard, J.C. Gibeaut, and A.L.Neuenschwander, "Fusion of airborne polarimetric and interferometric SAR data for classification of coastal environments," *IEEE Trans. Geosci. Rem. Sens*, **37**(3): 1306-1315, 1999.

[26] T. G. Dietterich and R. Bakiri, "Solving multiclass learning problems using error correcting output codes, *J. Artificial Intelligence Research*, **2**(1): 263-286, 1995.

[27] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, 1984.

[28] G. B. Dattatreya and L.N. Kanal, "Decision Trees in Pattern Recognition," in L.N. Kanal, A. Rosenfeld (eds.), *Progress in Pattern Recognition 2*, Elsevier, North-Holland, Amsterdam, 1985.

[29] S. Kumar, J. Ghosh, and M.M. Crawford, "Hierarchical Fusion of Multiple Classifiers for Hyperspectral Data Analysis," *Int. J. Pattern Analysis and Applications*, **5**(2): 210-220, 2002.

[30] M. Crawford, J. Ham, and J. Ghosh, "Robust classifiers for hyperspectral data analysis using limited training data," presented at the *2003 Tyrrhenian International Workshop on Remote Sensing*, Elba Island, Italy, Sept. 15-18.

[31] A.L. Neuenschwander, M.M. Crawford, and S. Ringrose, "Results of the EO-1 experiment - Use of Earth Observing-1 Advanced Land Imager (ALI) data to assess the vegetational response to flooding in the Okavango Delta, Botswana," *Int.. J. Remote Sens.*, accepted for publication.

[32] N.C. Oza and K. Tumer, "Input decimation ensembles: decorrelation through dimensionality reduction," *Proc. Int. Joint Conf. on Neural Networks*, Wahsington, D.C., 1999.

ACKNOWLEDGEMENTS

This research was supported by the NASA EO-1 Program (Grant NCC5-463), the Terrestrial Sciences Program of the Army Research Office (DAAG55-98-1-0287) and NSF (Grant IIS-0312471). We thank Amy Neuenschwander of the UT Center for Space Research for help in pre-processing the Hyperion data and interpreting the overall classification results.

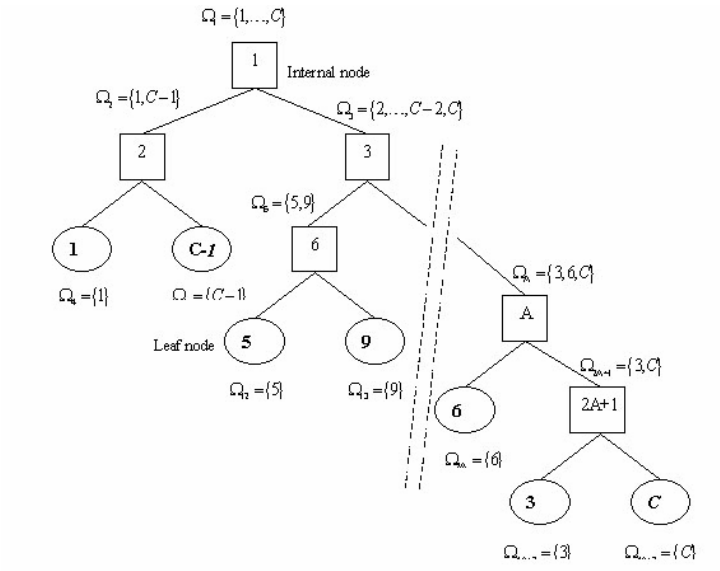

Figure 1. BINARY HIERACHICAL(multi)-CLASSIFIER framework for solving a *C*-class problem.

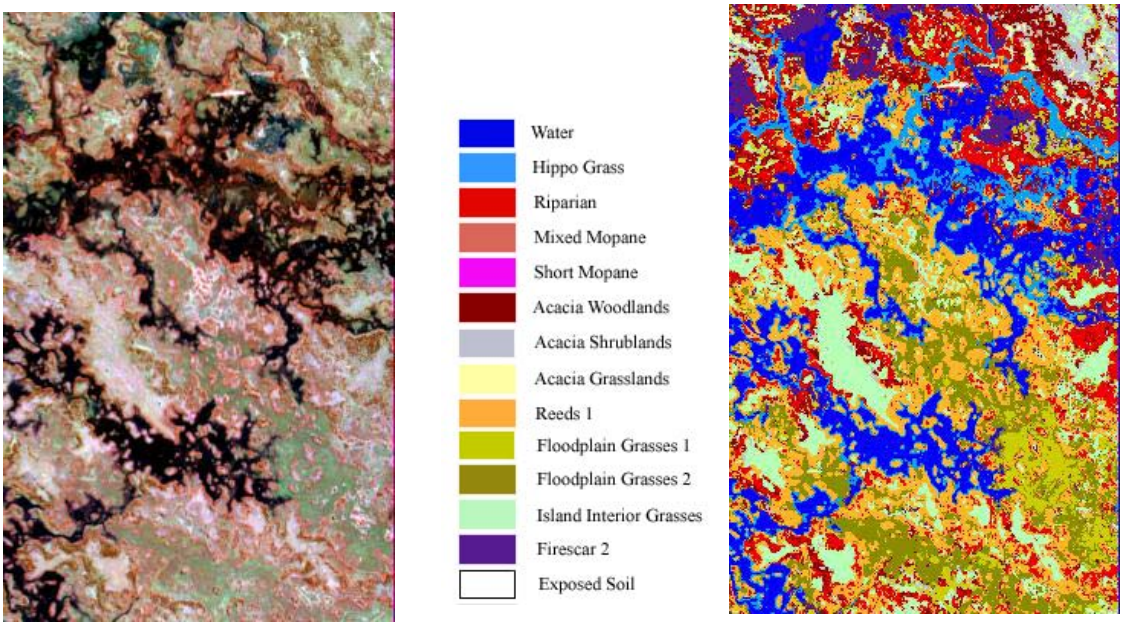

Figure 2 a) Subset of Hyperion data (Bands 51, 149, 31), b) Classified image of Hyperion data

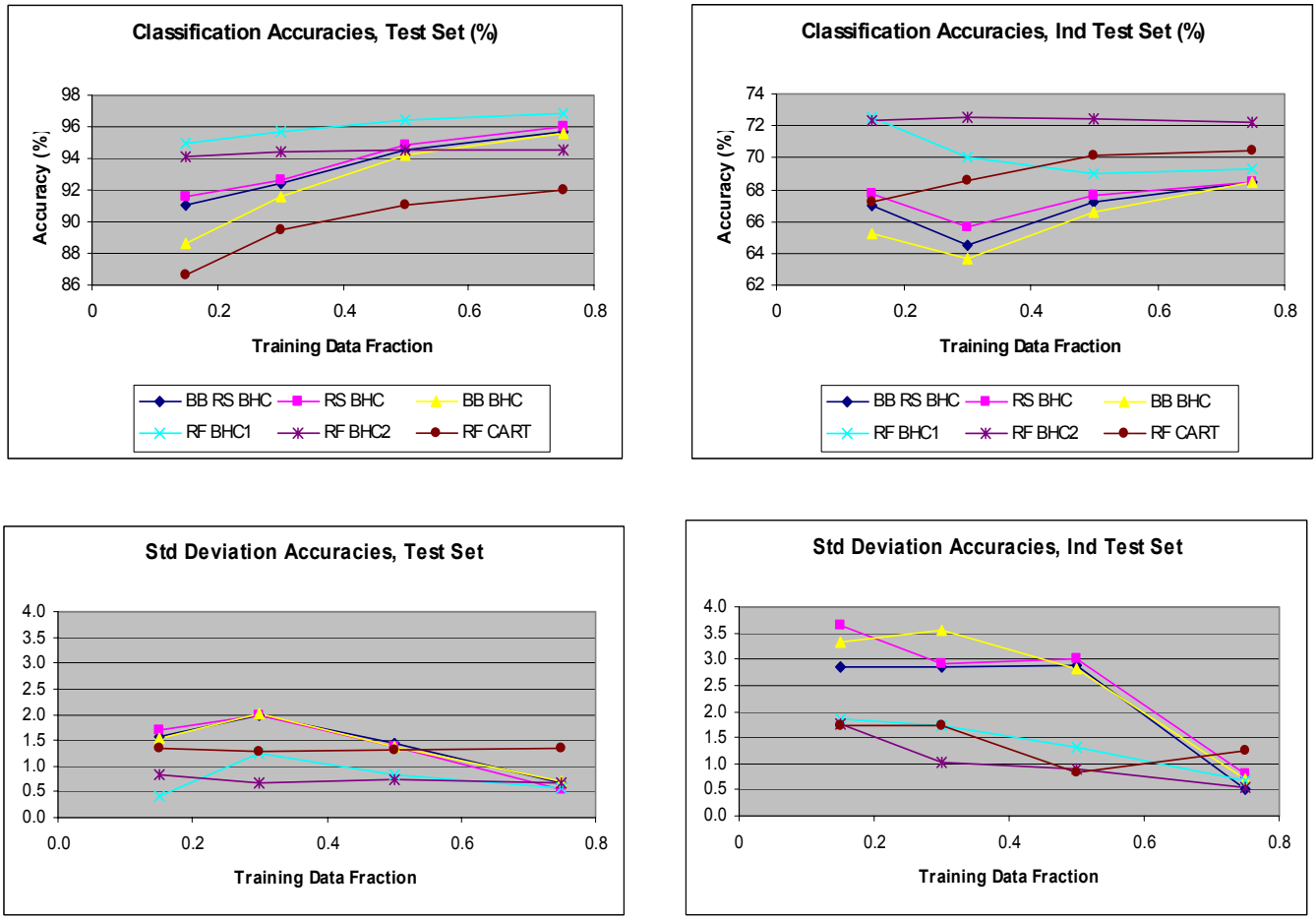

Figure 3. Average classification accuracies and standard deviations of accuracies for experiments on the test data and the independent test set using different fractions of training data.

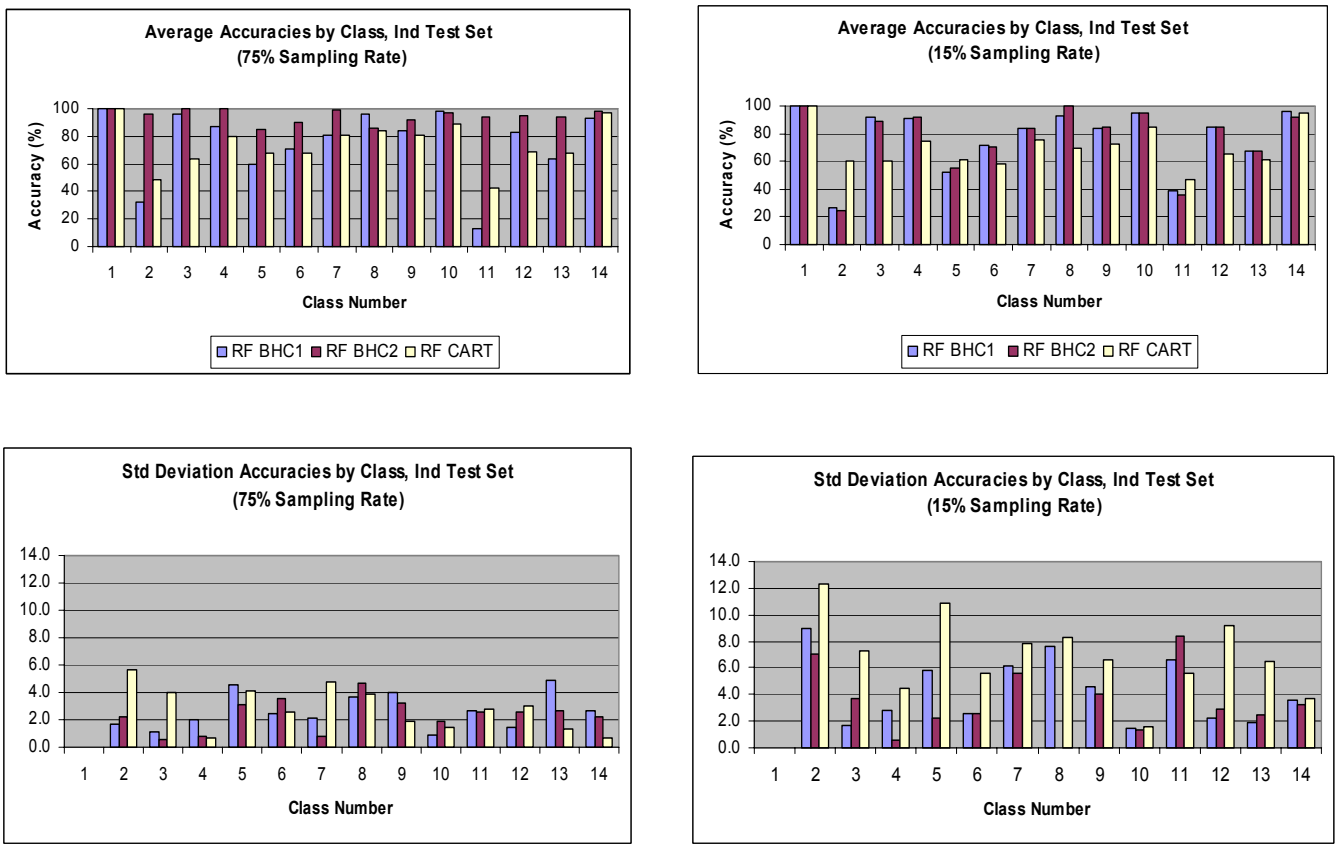

Figure 4. Average classification accuracies and standard deviations of accuracies by class for experiments performed using the 75% and 15% training fraction rates (as extremes).