

Adaptive feature spaces for land-cover classification with limited ground truth

Joseph T. Morgan¹, Alex Henneguelle², Melba M. Crawford¹, Joydeep Ghosh², and Amy Neuenschwander¹

¹Center for Space Research
{JMorgan, Crawford, Amy}@csr.utexas.edu
²Department of Electrical and Computer Engineering
The University of Texas at Austin
{Hennegue, Ghosh}@ece.utexas.edu

Abstract. Classification of hyperspectral data is challenging because of high dimensionality ($O(100)$) inputs, several possible output classes with uneven priors, and scarcity of labeled information. In an earlier work, a multiclassifier system arranged as a binary hierarchy was developed to group classes for easier and progressive discrimination [27]. This paper substantially expands the scope of such a system by integrating a feature reduction scheme that adaptively adjusts to the amount of labeled data available, while exploiting the highly correlated nature of certain adjacent hyperspectral bands. The resulting best-basis binary hierarchical classifier (BB-BHC) family is thus able to address the “small sample size” problem, as evidenced by our experimental results.

1 INTRODUCTION

The increasing availability of data from hyperspectral sensors has generated tremendous interest in the remote sensing community because they characterize the response of targets (spectral signatures) with greater detail than traditional sensors and thereby can improve discrimination between targets [7, 28]. A common application is to determine the land cover label of each (vector) pixel using supervised classification in which labeled training data or “ground truth”, \mathbf{X} , are used to estimate the label-conditional probability density functions, $P(x_1, x_2, \dots, x_D | L_i), i = 1, \dots, C$, or to directly estimate the a posteriori class probabilities. Unfortunately, hyperspectral data comes with some challenges as well. The dimensionality of the data (D) is high (~ 200) and the number of classes C is often 10 or more. To counter both these issues, Kumar et al. proposed in MCS 2000 [27] a method to decompose a ($C > 2$)-class problem into a binary hierarchy of ($C - 1$) simpler 2-class problems, that could be solved using a corresponding hierarchy of classifiers, each using a simple discriminant (Fisher projection). This top-down Binary Hierarchical Classifier (TD-BHC) provided superior accuracy and also yielded valuable domain knowledge.

This paper addresses a different challenge arising from the scarcity of labeled data, which is often of limited quantity in relation to the dimensionality D , at least for some of the poorly represented classes. This leads to well studied “small sample size” problems. For example, a classifier using Fisher’s linear discriminant function requires the inversion of the within-class covariance matrix. For the covariance matrix of D -dimensional data, there are $D(D+1)/2$ parameters to estimate and, at a minimum, there needs to be $D+1$ observations in each class to ensure a non-singular/invertible class specific covariance matrices [1]. Preferably, one should have at least $5D$ data points/class for good covariance estimation. Existing hyperspectral classifiers including the BHC are thus susceptible to small sample size issues [2].

This paper introduces a hierarchical, multiclassifier method that applies to multiclass problems while addressing the issue of limited training sets. It exploits both regularization of covariance estimates and adaptive feature reduction exploiting domain knowledge. Experiments show that it provides substantial improvements over the BHC when data is scarce without compromising performance for larger data sets.

2 PREVIOUS WORK

2.1 Small sample size problems

The substantial body of work in this area can be largely categorized into one of three different approaches [8]. Regularization methods including “shrinkage” try to stabilize the estimated covariance matrix directly by weighting the sample covariance matrix as well as “supplemental” matrices [17]. The covariance matrix can be “shrunk” towards the identity matrix or a pooled covariance matrix and there are hybrids that give weights to sample covariance (normal and diagonal) and a pooled covariance (normal and diagonal) matrix [5, 14]. While this may reduce the variance of the parameter estimates, the bias of the parameter estimates may increase dramatically. Rather than stabilize the covariance matrix directly, the pseudo-inverse of the covariance matrix can be used in place of the true inverse. Pseudo-inversion utilizes the non-zero eigenvalues of the covariance matrix [15, 17]. However, in addition to poor performance when the ratio of training data to dimensionality is very small, the pseudo-inverse has a “peaking effect” in its performance. It has been shown that the pseudo-inverse performs best when $|\mathbf{X}| = D/2$ and that the performance degrades as $|\mathbf{X}|$ approaches D [12, 16].

An alternate approach is to transform the input space into a reduced feature space via feature extraction or selection [10, 15], or to artificially add labeled samples. The transformations may result in some loss of interpretability, and may be poorly estimated due to the limited data. Specific techniques for identifying and augmenting the existing training data with unlabeled data already exist and have been shown to enhance strictly supervised classification [3, 9, 18-23]. However, not only can convergence of the updating scheme could be problematic, but also it is affected by selection of the initial training samples and by outliers.

A third approach is to use an ensemble of “weaker” classifiers. Bagging, Simple Random Sub-sampling, and Arcing involve taking subset samples for the original data and generating a classifier specific to each sub-sample [13]. When the data set is very small however, these methods have problems as well since the degradation in individual classifier performance (because of lack of data) cannot be compensated for by the gains from using an ensemble [29].

2.2 Hyperspectral Classification.

Among previous work on hyperspectral classification, the most relevant one here is the top down BINARY HIERARCHICAL CLASSIFIER (TD-BHC) framework that creates a multiclassifier system with C-1 classifiers arranged as a binary tree [27]. 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 $(\Omega_{2n}, \Omega_{2n+1})$, till at the leaves one obtains the original C classes [4]. The tree structure allows the more natural and easier discriminations to be accomplished earlier [6]. Subsequently, a bottom-up version (BU-BHC) was developed based upon an agglomerative clustering algorithm applied to merging the two most “similar” meta-classes until only one meta-class remains. Fisher’s discriminant is again used as the distance measure for determining the order in which the classes are merged. These two algorithms provided superior results when compared with a variety of other approaches to classifying hyperspectral data. Figure 1 depicts an example of a C-class BHC.

2.3 Hyperspectral Feature Reduction

From the domain knowledge in this field, it is known that the original input features - the bands of the hyperspectral data - that are “spectrally close” to one another, tend to be highly correlated. Jia and Richards proposed a Segmented Principal Components Transformation (SPCT) that exploits this characteristic [25, 26]. Edge detection algorithms were used to transform the original D individual bands into subsets of adjacent bands that are highly correlated based upon the estimated population correlation

matrix. From each subset, the most significant principal components are selected to give a feature vector that is significantly smaller than D . Although this approach utilizes high correlation between adjacent bands in hyperspectral data, it does not guarantee good discrimination capability because PCT aims at preserving variance in the data rather than maximizing discrimination among classes. Additionally, the segmentation approach of SPCT is based upon the correlation matrix over all of classes, and thus loses the often significant variance in the class conditional correlation matrices. Subsequently Kumar *et al.* proposed band combination techniques inspired by Best Basis functions [7]. Adjacent bands were selected for merging (alt. Splitting) in a bottom-up (alt. Top down) fashion using the product of a correlation measure and a Fisher based discrimination measure [4]. Although these two methods utilize the ordering of the bands and yield excellent discrimination, they are computationally expensive. Additionally, the quality of the discrimination functions, and thus the structure of the resulting feature space, will be affected by the amount of training data and this critical issue is not addressed.

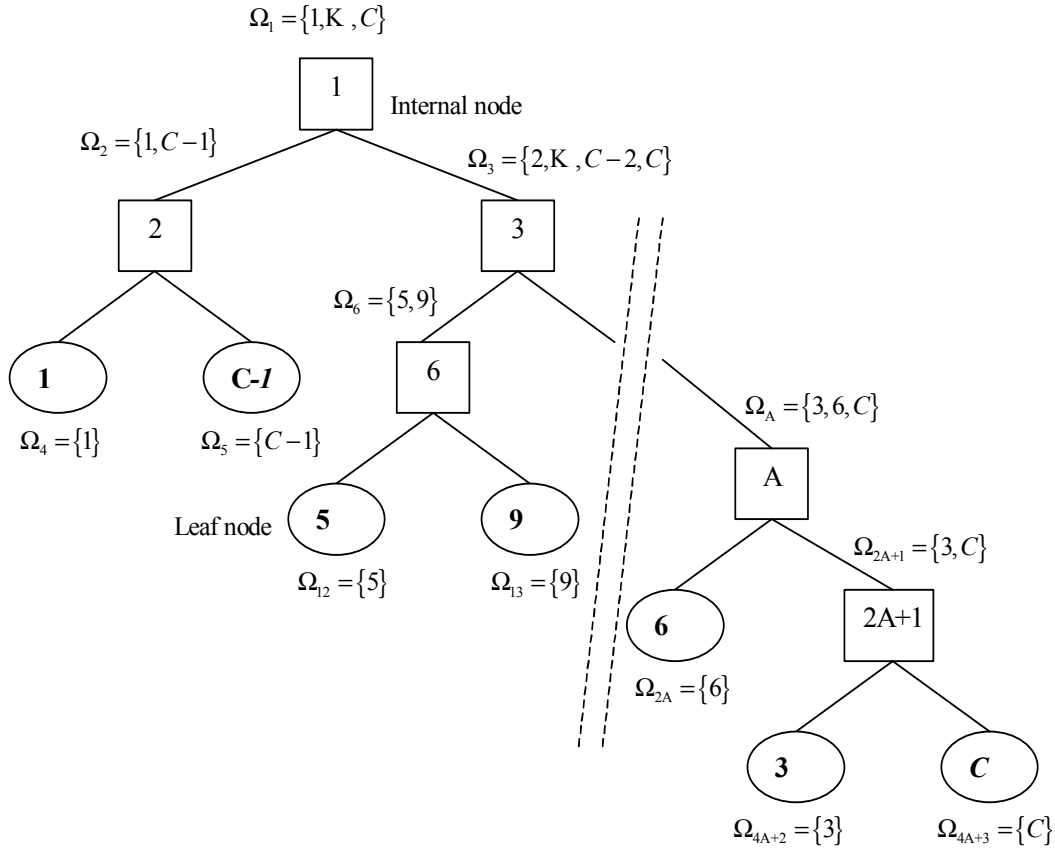


Figure 1. An example of a BINARY HIERARCHICAL(multi)-CLASSIFIER for solving a C -class problem. Each internal node n comprises of a feature extractor, a classifier, a left child $2n$, and a right child $2n+1$. Each node n is associated with a meta-class Ω_n .

3. Adaptive Best-Basis Binary Hierarchical Classifiers

The proposed approach applies a best-basis band-combining algorithm in conjunction with the BHC framework, while tuning the amount of feature reduction to the amount of data available. It also exploits the discovered hierarchy of classes to regularize covariance estimates using shrinkage.

3.1 Integrating Band Combination into Hierarchical, Multi-Classier Systems

The proposed approach can be viewed as a “best-basis” version of BHC (BB-BHC) that performs a band-combining algorithm prior to the partitioning (top down variant) or combining (bottom-up variant) of meta-classes. Band combination is done on highly correlated AND spectrally adjacent bands as this intuitively leads to the least loss in discrimination power. Because the correlation between bands varies among classes, the band reduction algorithm must be class dependent. In order to estimate the “correlation” for a group of bands (meta-bands) $B = [p : q]$ over a set of classes Ω , we define the correlation measure $Q(B)$ as the minimum of all the correlations within that group:

$$Q(B) = \min_{L_k \in \Omega} \min_{p \leq i < j \leq q} Q_{i,j}^{L_k} = \min_{L_k \in \Omega} \min_{p \leq i < j \leq q} \frac{S_{i,j}^{L_k}}{\sqrt{S_{i,i}^{L_k} S_{j,j}^{L_k}}} \quad (1)$$

where $S_{i,j}^{L_k}$ is the (i, j) th element of the sample covariance matrix for class L_k . The correlation measure (1) is used to determine which set of adjacent meta-bands should be merged at each successive step of the algorithm. Once the number of group bands is small enough, we maximize the discrimination between classes in the reduced space.

To address small sample sizes, rather than using a threshold on the correlation measure to determine if bands or group-bands should be merged, our algorithm focuses on preserving as many of the original bands as possible, commensurate with the amount of training data available. Thus the band-combining algorithm ensures that the least amount of discriminatory information is lost while trying to achieve a satisfactory ratio of training data to dimensionality. Because literature recommends different thresholds for the minimum $\alpha_{\text{ratio}} \leq \frac{|X|}{D}$, we allow this to be a user-defined input. Note that $|X|$ represents the number of data points in a child meta-class, and this number decreases as we go towards the leaves.

In pseudo-code, the adaptive band-combining algorithm that is performed before partitioning or merging meta-classes is:

1. $D^* = \min \left(D, \frac{|X|}{\alpha_{\text{ratio}}} \right)$
2. Initialize $l = 0$, $N_0 = D$, and $B_l^k = [k : k]$, $\forall k = 1, \dots, D$
3. If $N_l > D^*$ then continue. Otherwise, stop.
4. Find the best pair of band to merge: $K = \arg \max_{k=1, \dots, N_l-1} Q(B_l^k \cup B_l^{k+1})$
5. Update band structure:
 - $l = l + 1$, $N_l = N_{l-1} - 1$
 - If $K > 1$ then $B_l^k = B_{l-1}^k$, $\forall k = 1, \dots, K - 1$
 - $B_l^K = B_{l-1}^K \cup B_{l-1}^{K+1}$
 - If $K < N_l$ then $B_l^k = B_{l-1}^{k+1}$, $\forall k = K + 1, \dots, N_l$
6. Return to step 3.

3.2 Best Basis and Limited Data

When constructing a basis specific to each split in the BB-BHC, the quality of the correlation measure, computed from the class condition covariance matrices, will be dependent on the quantity of training data available to estimate the meta-class covariance matrices. This will become even more relevant for the “low branches” of the BB-BHC as the meta-classes become smaller in cardinality and the amount of training

data is strictly decreasing. In particular, the class specific correlation matrices $Q_{i,j}^{L_k} = \frac{S_{i,j}^{L_k}}{\sqrt{S_{i,i}^{L_k} S_{j,j}^{L_k}}}$ are

required in (1) to estimate the correlation measure $Q(B)$. However, if the label specific S^{L_k} covariance matrices are not suitable for inversion, failure to stabilize their estimation before constructing the basis unsatisfactorily passes the disadvantage of the small sample size from the estimate of Fisher's discriminant and linear discriminant function to the basis construction. Therefore, the label specific sample covariance matrices need to be stabilized. We define the ancestor sample covariance matrix S^{Anc} as being the sample covariance matrix which is estimated from at least $\alpha_{\text{ratio}} |X|$ observations and is most closely related to L_k based upon the BB-BHC structure. Because the trees are constructed in top-down and bottom-up manners, the search for S^{Anc} is performed uniquely for each type. In the top-down framework, if meta-class Ω_k is being considered for partitioning, then $S^{\Omega_k} = \sum_{L_i \in \Omega_k} P(L_i) S^{L_i}$ is the first candidate for S^{Anc} . However,

if $|X_{\Omega_k}| < \alpha_{\text{ratio}} D$, then the BB-BHC tree structure is climbed in search of a meta-class where $|X_{\Omega_k}| \geq \alpha_{\text{ratio}} D$. With the bottom-up framework, if $\{\Omega_{2n}, \Omega_{2n+1}\}$ are being considered for agglomeration, the first candidate for S^{Anc} is $S^{\text{Pooled}} = P(\Omega_{2n}) S^{\Omega_{2n}} + P(\Omega_{2n+1}) S^{\Omega_{2n+1}}$. However, because the BB-BHC is being constructed bottom-up, the structure cannot be climbed in search of a suitable S^{Anc} . Therefore, if $|X_{\Omega_i + \Omega_j}| < \alpha_{\text{ratio}} D$, then $S^{\text{Anc}} = \sum_{i=1}^C P(L_i) S^{L_i}$. Note that this estimate for

S^{Anc} will be used, even when the total quantity of training data available is less than $\alpha_{\text{ratio}} D$. When applicable, the stabilized estimates of the label specific covariance matrices are used to estimate the correlation measure (1).

4 RESULTS

Evaluation of the proposed D-BB-BHC algorithm was performed on two separate sites: the Bolivar Peninsula, located at the mouth of Galveston Bay, Texas and NASA's John F. Kennedy Space Center (KSC) at Cape Canaveral, Florida.

4.1 Bolivar Peninsula

Bolivar Peninsula is located at the mouth of Galveston Bay and is part of the low relief barrier island system on the Texas Gulf coast. The area contains two general vegetation types, wetlands and uplands, with the marsh area further characterized in terms of sub-environments defined by the wetland maps. For classification purposes, 11 classes representing the various land cover types that occur in this environment have been defined for the site (Table 1). HyMap (Hyperspectral Mapper) was used to acquire data over Bolivar Peninsula on September 17, 1999, at a spatial resolution of 5m. HyMap, an airborne hyperspectral optical sensor developed in Australia, acquired the data in 126 bands with almost contiguous spectral coverage over the wavelength range of 0.44-2.48 [24]. For this particular acquisition, only four bands {63, 64, 95, 126} were dominated by water absorption, resulting in a low signal to noise ratio, and therefore not considered in subsequent analysis. In this case, the practical dimensionality D is 122.

Multiple experiments were accomplished on this site using stratified (class specific) sampling at percentages of: 75, 50, 30, 15, 5, and 1.5. It is interesting to note that even at the sampling percentage of 75, the amounts of training data for classes 5 and 10 are still less than D (sand flats $|X_{L_5}| = 86$ and

Table 1. Classes for Bolivar Peninsula and the quantity of training data per class

Class	Name	Total Obs
1	Water	1019
2	Low Proximal Marsh	1127
3	High Proximal Marsh	910
4	High Distal Marsh	752
5	Sand Flats	148
6	Ag 1 (pasture)	3073
7	Trees	222
8	General Uplands	704
9	Ag 2 (bare soil)	1095
10	Transition Zone	114
11	Pure Silicornia	214

transition zone $|\mathbf{X}_{L_{10}}| = 111$). We used $\alpha_{\text{ratio}} = 5$ for all sampling percentages except for 1.5 ($\alpha_{\text{ratio}} = 1.5$).

The lower threshold was used to ensure that there were at least two observations per label L_i . Ten experiments, using simple random sampling, were performed at each percentage for the bottom-up and top-down frameworks of the traditional BHC [TD-BHC, BU-BHC], the traditional BHC using the pseudo-inverse for tree construction (estimating Fisher’s discriminant) and feature extraction (calculating Fisher’s linear discriminant function), [TD-P-BHC, BU-P-BHC], and the adaptive best-basis BHC [TD-BB-BHC, BU-BB-BHC]. The results are presented in Figure 2.

By adapting the size of the feature space to reflect the amount of training data available, a high level of classification accuracy is preserved for extremely low number of observations. At the 50th percentage of

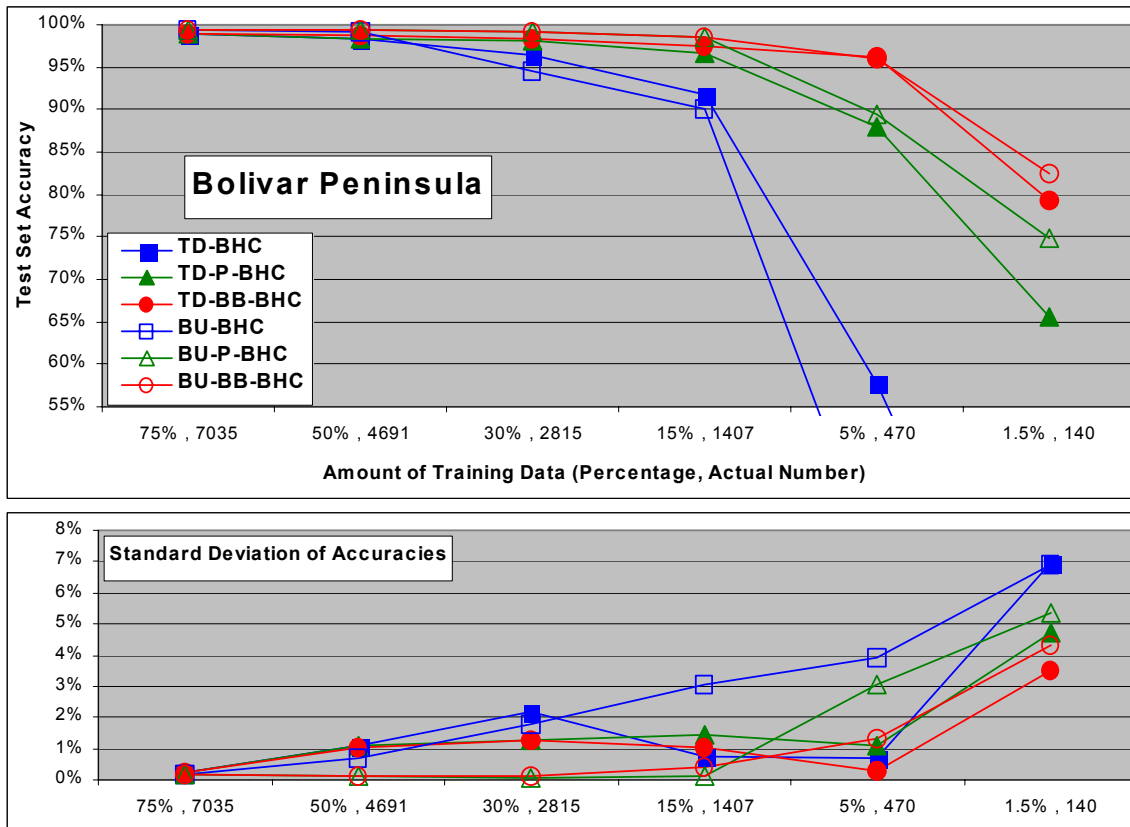


Figure 2. Classification (test set) accuracies for Bolivar Peninsula

sampling, which is typically used to separate data sets into training and testing, the BB-BHC actually performs slightly better than the BHC. Importantly, even though at the 50th percentage using the pseudo-inverse does not improve the results because there are at least $D+1$ observations per L_i , the results indicate that even though the covariance matrices are non-singular, they are still poorly estimated and will result in poor inverted matrices. Not only does the BB-BHC perform the best at every sampling percentage relative to the other TD and BU classifiers, but also the accuracies are generally more stable (smaller standard deviation of accuracies). This is important because different investigators may vary in terms of the areas they consider as ground truth. Combating the limited training data by using the correlation matrix for feature reduction helps retain the information necessary for successful land-cover prediction. Classification accuracies of over 80% can still be achieved even with only 140 total labeled samples and only 2 labeled pixels available for classes 5 (sand flats) and 10 (transition zone).

4.2 Cape Canaveral

The wetlands of the Indian River Lagoon system, located on the western coast of the Kennedy Space Center (KSC) at Cape Canaveral, Florida, are a critical habitat for several species of waterfowl and aquatic life. The test site for this research consists of a series of impounded estuarine wetlands of the northern Indian River Lagoon (IRL) that reside on the western shore of the Kennedy Space Center. Classification of land cover for this environment is difficult due to the similarity of spectral signatures for certain vegetation types. For classification purposes, 13 classes representing the various land cover types that occur in this environment have been defined for the site (Table 2).

Table 2. Classes for KSC and the quantity of training data per class

Class	Name	Total Obs
1	Scrub	761
2	Willow Swamp	243
3	CP Hammock	256
4	CP/Oak Hammock	252
5	Slash Pine	161
6	Oak/Broadleaf Hammock	229
7	Hardwood Swamp	105
8	Graminoid Marsh	420
9	Spartina Marsh	520
10	Cattail Marsh	397
11	Salt Marsh	419
12	Mud Flats	447
13	Water	927

The NASA AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) spectrometer was used to acquire data over the Kennedy Space Center, Florida on March 23, 1996. AVIRIS acquires data in 224 bands of 10 nm widths in the reflected visible and near infrared spectrum (400 - 2500 nm). The data, acquired from an altitude of approximately 20km, have a spatial resolution of 18 m [19]. Unfortunately, forty-eight bands collected by the AVIRIS sensor are dominated by water absorption, which results in a low signal noise ratio, and are not considered in subsequent analysis. In this case, $D=176$ bands of AVIRIS data are used {bands 1-4, 102-116, 151-172, and 218-224 have been removed}. Multiple experiments were accomplished on this site using stratified (class specific) sampling at percentages of: 75, 50, 30, 15, 5, and 1.5. At the sampling percentage of 75, the amounts of training data for classes 5, 6, and 7 are still less than D and, at the 50th percentage, so are classes 2, 3, and 4. Ten experiments, using simple random sampling, were performed at each percentage for the bottom-up and top-down frameworks of the traditional BHC [TD-BHC, BU-BHC], the traditional BHC using the pseudo-inverse for tree construction (estimating Fisher's discriminant) and feature extraction (calculating Fisher's linear discriminant function), [TD-P-

BHC, BU-P-BHC], and the adaptive best-basis BHC [TD-BB-BHC, BU-BB-BHC]. The results are presented in Figure 3.

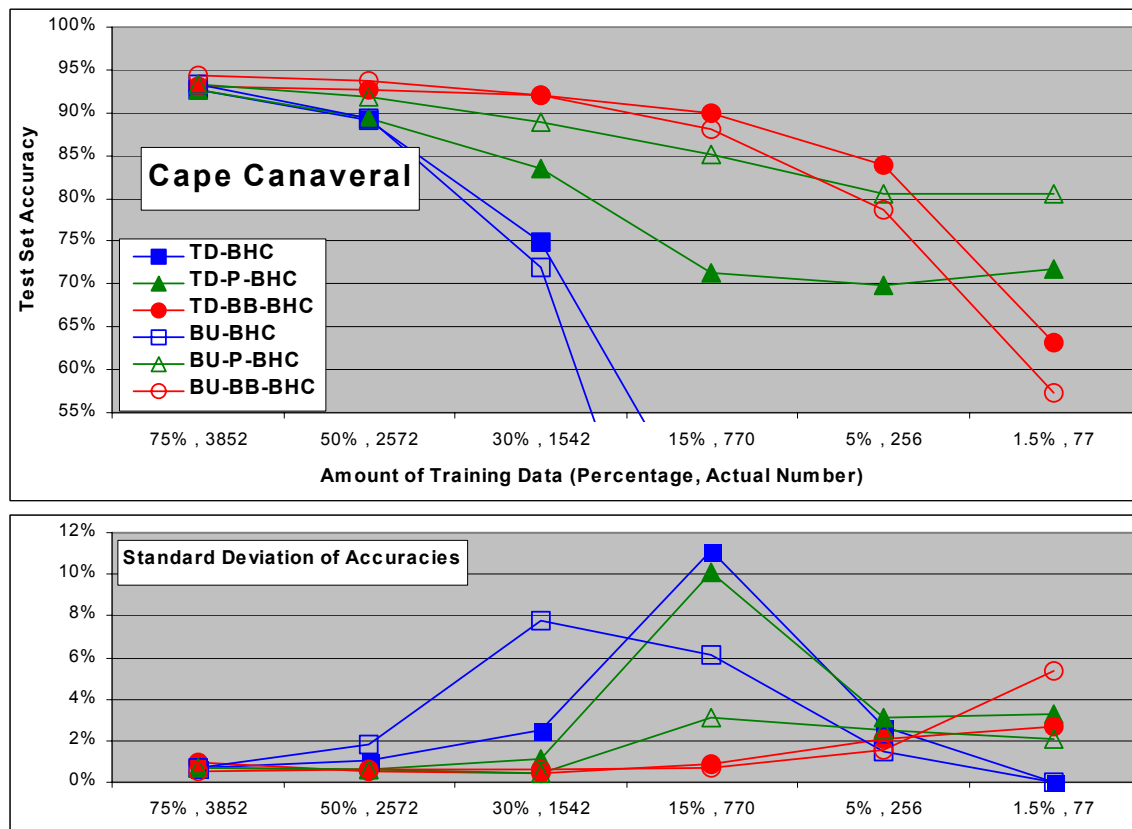


Figure 3. Classification (test set) accuracies for Cape Canaveral

The test set accuracies for Cape Canaveral are very similar to those of Bolivar Peninsula except that the pseudo-inverse classifiers perform better at the 1.5% sampling rate, with the accuracies for the pseudo-inverse BHC classifiers maintaining the accuracy level that had been achieved at the 5% sampling rate. At the lower sampling percentages, the covariance matrices are very poorly estimated in the full dimensional space, yet the accuracies are still fairly high using pseudo-inversion indicating that the differences in class means is the main reason the level of discrimination is being maintained. This result is also reflected by the standard deviations of the accuracies, which spike in the 15%-30% sampling rate range for the pseudo-inverse classifiers where the covariance matrices are still helping maintain a higher level of classification accuracy (than the 1.5%-5% range), though unstable. Also, an explanation for diminished classification accuracies of the BB-BHC at the 1.5% sampling rate is that there might be a minimum requirement, the “intrinsic dimensionality” [11], for the number of bands, after which the results drop off sharply.

5 CONCLUSIONS AND FUTURE WORK

The dependency of classification accuracy upon the ratio of training data size to the dimensionality of the data has been widely noted and needs to be addressed during the design of a classifier. While the advent of hyperspectral sensors has provided unique opportunities in remote sensing, the increased dimensionality of the data necessitates that researchers pay even more attention to designing classifiers that are more tolerant of the quantity of training data available. This paper proposed a multi-classifier framework that utilizes the flexibility gained by transforming the output space and input space simultaneously to combat the small

sample size problem. By reducing the size of the feature space in a directed manner, dependent upon the quantity of training data available in the binary hierarchy of meta-classes, a high level of classification accuracy is preserved even when faced with low quantities of training data for some classes.

Combating the small sample size problem with the dynamic best-basis algorithm helps preserve the interpretability of the data, but using Fisher's linear discriminant function as the feature extractor at each internal node of the BHC diminishes this attractive characteristic. While the discriminant function weights on each band/group-band could be analyzed to determine the respective band's importance, the interpretation and insight would be less complicated if feature selection was performed rather than feature extraction. Therefore, using feature selection rather than feature extraction, and the likely trade-off between classification accuracy and retention of domain knowledge, should be investigated further.

References

1. T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons, 1984.
2. D. Landgrebe, "Information extraction principles and methods for multispectral and hyperspectral image data," *Information Processing for Remote Sensing*, ed. Chen, C.H., World Scientific Pub. Co, NJ, 1999.
3. S. Tadjudin and D.A. Landgrebe, "Robust parameter estimation for mixture model," *IEEE Trans on Geosci and RS*, vo. 38, no. 1, pp. 439-45, Jan. 2000.
4. S. Kumar, J. Ghosh and M. M. Crawford, "Hierarchical fusion of multiple classifiers for hyperspectral data analysis," *Pattern Analysis and Applications*, Special Issue on Classifier Fusion (to appear).
5. S. Tadjudin and D.A. Landgrebe, "Covariance estimation with limited training samples," *IEEE Trans on Geosci and RS*, vol. 37, no. 4, pp. 2113-8, 1999.
6. P.A. Devijver and J. Kittler (editors), *Pattern Recognition Theory and Application*. Springer-Verlag, 1987.
7. S. Kumar, J. Ghosh, and M.M. Crawford, "Best basis feature exaction algorithms for classification of hyperspectral data," *IEEE Trans on Geosci and RS*, vol. 39, issue 7, pp. 1368-79, July 2001.
8. S.J. Raudys and A.K. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners", *IEEE Trans on PAMI*, vol. 13, no. 3, pp. 252-64, March 1991.
9. Qiong Jackson and David Landgrebe, "An adaptive classifier design for high-dimensional data analysis with a limited training data set", *IEEE Trans on Geosci and RS*, vol. 39, issue 12, pp. 2664-79, Dec 2001.
10. T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons, 1984.
11. Andrew Webb, *Statistical pattern recognition*. London: Oxford University Press, 1999.
12. Marina Skurichina, "Stabilizing weak classifiers," Thesis, Vilnius State Univesity, 2001.
13. L. Breiman, "Bagging predictors," *Machine Learning*, 24(2): 123-40, 1996.
14. A. McCallum, R. Rosenfeld, T. Mitchell, and A.Y. Ng, "Improving text classification by shrinkage in a hierarchy of classes," Proc. 15th International Conf. on Machine, Madison, WI, Morgan Kaufmann, San Mateo, CA, pp. 359-67 1998.
15. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed, Boston, 1990.
16. Sarunas Raudys and Robert P. W. Duin, "Expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix," *Pattern Recognition Letters*, vol 19, pp 385-92, April 1998.
17. 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, pp. 891-6, 1996.

18. A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," *Proceedings of the 11th Annual Conf. Computational Learning Theory*, pp. 92-100, 1998.
19. Webpage. Jet Propulsion Lab, California Institute of Technology, <http://makalu.jpl.nasa.gov/>.
20. B. Jeon and D. Landgrebe, "Partially supervised classification using weighted unsupervised clustering," *IEEE Trans on Geosci and RS*, vol. 37, no. 2, pp. 1073-9, March 1999.
21. T.M. Mitchell, "The role of unlabeled data in supervised learning," *Proc. Sixth International Colloquium on Cognitive Science*, 8pgs, 1999.
22. V.R. de Sa, "Learning classification with unlabeled data," *Advances in Neural Information Processing Systems 6*, 1994.
23. 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 and RS*, vol. 32, no. 5, pp. 1087-95, 1994.
24. T. Cocks, R. Jenssen, A. Stewart, I. Wilson, and T. Shields, "The HyMap airborne hyperspectral sensor: the system, calibration and performance", Proc. 1st EARSeL Workshop on Imaging Spectroscopy (M. Schaepman, D. Schl pfer, and K.I. Itten, Eds.), Zurich, EARSeL, Paris, pp. 37-42, 6-8 October, 1998.
25. X. Jia, Classification Techniques for Hyperspectral Remote Sensing Image Data. PhD Thesis, Univ. College, ADFA, University of New South Wales, Australia, 1996.
26. X. Jia and J.A. Richards, "Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification", *IEEE Trans on Geosci and RS*, vol. 37, no. 1, pp. 538-42, 1999.
27. S. Kumar, J. Ghosh, and M. M. Crawford, "A hierarchical multiclassifier system for hyperspectral data analysis", First International Workshop on Multiple Classifier Systems, Sardinia, Italy, pp. 270-9, June 2000.
28. David Landgrebe, "Hyperspectral image data analysis as a high dimensional signal processing problem," (Invited), Special Issue of the *IEEE Signal Processing Magazine*, vol 19, no 1 pp. 17-28, January 2002.
29. K. Tumer and J. Ghosh, "Error correlation and error reduction in ensemble classifiers" *Connection Science*, Special Issue on Combining, Vol. 8, No. 3/4, Dec 1996, pp. 385-404