Best Bases Bayesian Hierarchical Classifier for Hyperspectral Data Analysis

Joseph T. Morgan[1], Alex Henneguelle[2], Melba M. Crawford[1], Joydeep Ghosh[2], and Amy Neuenschwander[1]
[1]Center for Space Research
{JMorgan, Crawford, Amy}@csr.utexas.edu
[2]Department 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 inputs coupled with possible high dimensional outputs and scarcity of labeled information. Previously, a multiclassifier system was formulated in a binary hierarchical framework to group classes for accurate, rapid discrimination. In order to improve performance for small sample sizes, a new approach was developed that utilizes a feature reduction scheme which adaptively adjusts to the amount of labeled data available, while exploiting the fact that certain adjacent hyperspectral bands are highly correlated. The resulting best-basis binary hierarchical classifier (BB-BHC) family is thus able to address the "small sample size" problem, as evidenced by experimental results obtained from analysis of AVIRIS and Hyperion data acquired over Kennedy Space Center.

INTRODUCTION

The increasing availability of data from hyperspectral sensors provides the capability to characterize the spectral response of targets with greater detail than multispectral sensors, and thereby can potentially improve discrimination between targets. However, the high dimensionality of the data is problematic for supervised statistical classification techniques that utilize the estimated covariance matrix since the number of known samples is typically small relative to the dimension of the data [1]. Previous research has dealt with this problem using a) regularization methods to stabilize the estimated covariance matrix directly or by using the pseudo-inverse [2,3], b) transformation of the input space via reduction in the dimension of the feature space via feature extraction or selection [4,5] or addition of artificially labeled data [6,7], and c) utilization of ensembles of classifiers (e.g. bagging, simple random sub-sampling, arcing) [8,9]. When sample sizes are very small, these approaches are inadequate. Regularized covariance matrices often produce biased estimates; the pseudo-inverse approach does not perform uniformly well over a range of sample sizes; feature extraction methods suffer from interpretability of results; and the performance of ensembles of classifiers is greatly degraded when sample sizes are extremely small. [10].

BEST-BASES BAYESIAN HIERARCHICAL CLASSIFIER

A new approach has been developed specifically to address the problem of extremely small sample sizes. It is based on a Binary Hierarchical Classifier (BHC) framework that creates a multiclassifier system with C-1 classifiers arranged as a binary tree [11]. In the top-down implementation, 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. The 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 leaf nodes [12]. The tree structure allows the more natural and easier discriminations to be accomplished earlier [13]. The bottom-up version of the BHC utilizes an agglomerative clustering algorithm whereby the two most "similar" meta-classes are merged 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. Both algorithms perform quite well for large dimensional input and output problems if data samples are not extremely small.

The new method extends the TD-BHC and BU-BHC approaches through a best bases feature extraction technique that exploits the highly correlated bands observed within hyperspectral data when it is advantageous. Jia and Richards proposed a Segmented Principal Components Transformation (SPCT) that also exploits this characteristic [14]. However, SPCT does not guarantee good discrimination capability because the PCT transformation

criterion is related to variance, not discrimination between classes. Further, the SPCT is based on the correlation matrix over all classes, and thus loses information from the class conditional correlation matrices. Kumar *et. al* proposed band combination techniques inspired by Best Basis functions [15]. Adjacent bands were selected for merging/splitting in a bottom-up/top down approach using the product of a correlation measure and a Fisher based discrimination measure. Although the methods exploit band ordering and yield excellent discrimination, they are computationally expensive. Additionally, the quality of the discrimination functions, and thus the structure of the resulting feature space, is affected by the amount of training data.

The new approach applies a best-basis band-combining algorithm in conjunction with the BHC framework, while tuning the amount of feature reduction to the quantity of available data. It also exploits the discovered hierarchy of classes to regularize covariance estimates using shrinkage. The method can be viewed as a "best-basis" BHC that performs a band-combining stage prior to the partitioning (TD variant) or combining (BU variant) of meta-classes. Band combination is performed on highly correlated, spectrally adjacent bands. Because the intraband correlation is class specific, 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, the correlation measure $Q(B)$ is defined as

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

where $S_{i,j}^{L_k}$ is the (*i, j*)th element of the sample covariance matrix for class L_k. The correlation measure is used to determine which set of adjacent meta-bands should be merged at successive steps of the algorithm. Once the number of group bands is small enough, discrimination between classes in the reduced space is maximized. When sample sizes are small, the algorithm focuses on preserving as many of the original bands as possible, commensurate with the amount of training data available. The minimum $\alpha_{\text{ratio}} \le \frac{|\boldsymbol{X}|}{D}$, this is a user-defined input.

When constructing a basis specific to each split in the BB-BHC, the quality of the correlation measure depends on the quantity of training data available to estimate the meta-class covariance matrices. This is particularly critical 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. However, if the label specific S^{L_k} covariance matrices are not suitable for inversion, failure to stabilize their estimates before constructing the basis passes the disadvantage of the small sample size from the estimate of Fisher's disciminant and linear discriminant function to the basis construction. The ancestor sample covariance matrix S^{Anc} is defined as the sample covariance matrix that is estimated from at least $\alpha_{\text{ratio}}|\boldsymbol{X}|$ observations and is most closely related to L_k based upon the BB-BHC structure. The search for S^{Anc} is performed uniquely for TD and BU structures. In the TD framework, if meta-class Ω_k is being considered for partitioning, than $S^{\Omega_k} = \sum_{L_i \in \Omega_k} P(L_i) S^{L_i}$ is the first candidate for S^{Anc}. However, if $\left|\boldsymbol{X}_{\Omega_k}\right| < \alpha_{\text{ratio}} D$, then the BB-BHC tree structure is climbed in search of a meta-class where $\left|\boldsymbol{X}_{\Omega_k}\right| \ge \alpha_{\text{ratio}} D$. With the bottom-up framework, if $\{\Omega_{2n}, \Omega_{2n+1}\}$ are being considered for agglomeration, the first candidate for S^{Anc} is $S^{\text{Pooled}} = P(\Omega_{2n}) S^{\Omega_{2n}} + P(\Omega_{2n+1}) S^{\Omega_{2n+1}}$. Because the BB-BHC is constructed bottom-up, the structure cannot be climbed in search of a suitable S^{Anc}. Therefore, if $\left|\boldsymbol{X}_{\Omega_i + \Omega_j}\right| < \alpha_{\text{ratio}} D$, then $S^{\text{Anc}} = \sum_{i=1}^{C} P(L_i) S^{L_i}$. (This estimate for S^{Anc} is 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.

RESULTS

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 that reside on the western

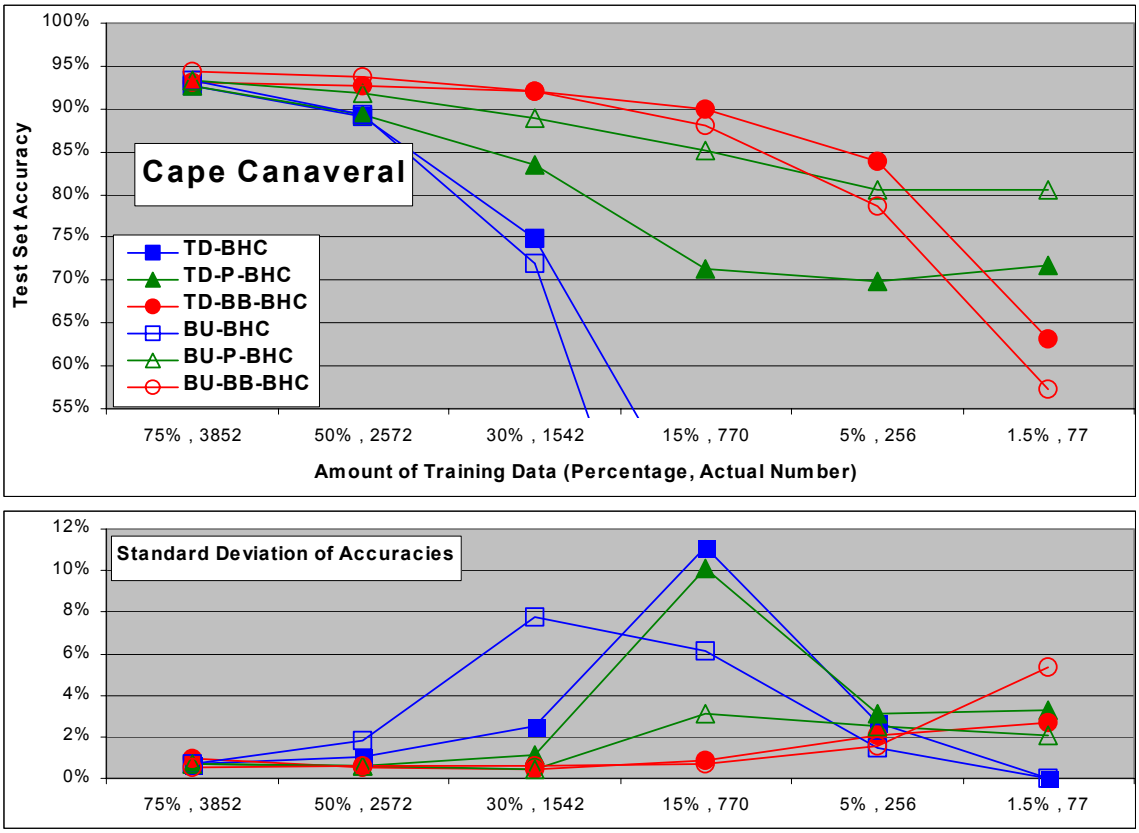

Figure 1. Classification (test set) accuracies for Cape Canaveral

shore of KSC. 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 and samples of indicated size collected: 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); Mud Flats (447); Water (927).

The NASA AVIRIS sensor acquired data over KSC on March 23, 1996. After removing water absorption bands, D=176 bands of data remained for classification. Multiple experiments were performed using stratified sampling at percentages of: 75, 50, 30, 15, 5, and 1.5. At 75% sampling, the amounts of training data for classes 5, 6, and 7 are still less then D, as are classes 2, 3, and 4 at 50%. Ten experiments, using simple random sampling, were performed at each percentage for the bottom-up and top-down frameworks of the traditional TD-BHC, BU-BHC, the TD-P-BHC and BU-P-BHC using the pseudo-inverse for tree construction and feature extraction, and the adaptive TD-BB-BHC, BU-BB-BHC best bases. Results are shown in Figure 1.

Both best bases BHC methods yield excellent results for sampling rates as low as 15%. Results are still acceptable for both the best bases and pseudo-inversion methods at the 5% sampling rate, although the TD version of the pseudo-inverse approach degrades more rapidly than the BU method. At the lower sampling percentages, the covariance matrices are very poorly estimated in the full dimensional space, yet test 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 although unstable. However, in general, diminished classification accuracies of the BB-BHC at the 1.5% sampling rate may be

due to a minimum requirement, the "intrinsic dimensionality", for the number of bands, after which the results degrade sharply [16]. Overall, results for KSC were consistent with those from experiments at other sites, with the exception of the results at the 1.5% sampling rate, where the BB-BHC often yielded better results than the pseudo-inverse.

CONCLUSIONS AND FUTURE WORK

A new best bases multi-classifier framework that utilizes the flexibility gained by transforming the output and input spaces simultaneously has been developed 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 instead of feature extraction. Thus, use of feature selection rather than feature extraction, and the likely trade-off between classification accuracy and retention of domain knowledge, is being investigated further. Additionally, alternative best bases approaches that incorporate approximations to the spectrum and an expanded feature space that includes estimates of derivatives are being developed and incorporated into this adaptive best bases framework.

REFERENCES

1. 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.
2. S. Tadjudin and D.A. Landgrebe, "Covariance estimation with limited training samples," *IEEE Trans. Geosci. Rem. Sens.*, 37(4), 2113-8, 1999.
3. 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.
4. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed, Boston, 1990.
5. M.M. Crawford, S. Kumar, M.R. Ricard, J.C. Gibeaut, and A.L. Neuenschwander, "Fusion of Airborne Polarimetric and Interfermetric SAR Data for Classification of Coastal Environments," *IEEE Trans. on Geoscience and Remote Sensing*, 37(4), 1306-1315, 1999.
6. Qiong 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.
7. 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.
8. Marina Skurichina, "Stabilizing weak classifiers," Thesis, Vilnius State Univesity, 2001.
9. L. Breiman, "Bagging predictors," *Machine Learning*, 24(2), 123-40, 1996.
10. 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.
11. S. Kumar, J. Ghosh, and M.M. Crawford, "A Hierarchical Multiclassifier System for Hyperspectral Data Analysis, *Lecture Notes in Computer Science*, 1857:270-279, 2000.
12. 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).
13. P.A. Devijver and J. Kittler (editors), *Pattern Recognition Theory and Application*. Springer-Verlag, 1987.
14. 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.
15 S. Kumar, J. Ghosh, and M.M. Crawford, "Best basis feature exaction algorithms for classification of hyperspectral data," *IEEE Trans. Geosci. Rem. Sens*., 39(7), 1368-79, 2001.
16. Andrew Webb, *Statistical pattern recognition.* London: Oxford University Press, 1999.