International Journal of Pattern Recognition and Artificial Intelligence
© World Scientific Publishing Company

ADAPTIVE FEATURE SPACES FOR LAND COVER CLASSIFICATION WITH LIMITED GROUND TRUTH DATA

JOSEPH T. MORGAN[1], ALEX HENNEGUELLE[2], JISOO HAM[1], JOYDEEP GHOSH[2] and MELBA M. CRAWFORD[1]

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

Classification of land cover based on hyperspectral data is very challenging because typically tens of classes with uneven priors are involved, the inputs are high dimensional, and there is often scarcity of labeled data. Several researchers have observed that it is often preferable to decompose a multi-class problem into multiple two-class problems, solve each such sub-problem using a suitable binary classifier, and then combine the outputs of this collection of classifiers in a suitable manner to obtain the answer to the original multi-class problem. This approach is taken by the popular error correcting output codes (ECOC) technique, as well by the binary hierarchical classifier (BHC). Classical techniques for dealing with small sample sizes include regularization of covariance matrices and feature reduction. In this paper we address the twin problems of small sample sizes and multi-class settings by proposing a feature reduction scheme that adaptively adjusts to the amount of labeled data available. This scheme can be used in conjunction with ECOC and the BHC, as well as other approaches such as round-robin classification that decompose a multi-class problem into a number of two (meta)-class problems. In particular, we develop the best-basis binary hierarchical classifier (BB-BHC) and best basis ECOC (BB-ECOC) families of models that are adapted to "small sample size" situations. Currently, there are few studies that compare the efficacy of different approaches to multi-class problems in general settings as well as in the specific context of small sample sizes. Our experiments on two sets of remote sensing data show that both BB-BHC and BB-ECOC methods are superior to their non-adaptive versions when faced with limited data, with the BB-BHC showing a slight edge in terms of classification accuracy as well as interpretability.

Keywords: multi-class problems, multiple classifier systems, hierarchical classifiers, error correcting output codes, small sample size problem, remote sensing.

1. Introduction

The increasing availability of data from hyperspectral sensors has generated tremendous interest in the remote sensing community because these instruments characterize the response of targets (spectral signatures) with greater detail than traditional sensors and thereby can potentially improve discrimination between targets[31,33]. A common application is to determine the land cover label of each (vector) pixel by

using labeled training data (ground truth), X, to estimate the parameters of the label-conditional probability density functions, $P(x_1, x_2, \ldots, x_D|L_i)$, $i = 1, \ldots, C$, or to directly estimate the aposteriori class probabilities. Unfortunately, classification of hyperspectral data is challenging for several reasons. The dimensionality of the data (D) is high (~ 200), and the number of classes C is often in the teens. The sensor measurements obtained from a given land cover type can vary somewhat over time and space, and thus the class-conditional likelihoods can vary from image to image. Obtaining labeled data is expensive and time consuming because it either involves field campaigns or manual interpretation of high resolution imagery. However, hyperspectral data tend to be correlated both spectrally and spatially, and these two properties can often be exploited to make the classification problem more tractable.

In our previous work on land cover classification[29], we had addressed the problem of being faced with a moderately large number of classes by systematically decomposing a C-class problem into a binary hierarchy of $C - 1$ simpler two-class problems that could be solved using a corresponding hierarchy of classifiers, each involving a simple discriminant (Fisher projection). The use of a simple feature extraction process also helped deal with the high dimensionality of the input space. The resulting top-down Binary Hierarchical Classifier (TD-BHC), which is really an ensemble of classifiers arranged as a hierarchy, provided superior results in terms of test accuracies as compared to using a variety of direct approaches to the multi-class problem. In addition, the hierarchies of classes automatically derived from the data yield valuable domain knowledge about the relationships among different types of land cover.

This paper addresses a different challenge stemming from the scarcity of labeled data, which is often of limited quantity relative to the dimensionality D, at least for some poorly represented classes. This leads to the well-studied small sample size problem. 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, minimally there must be $D + 1$ observations of each class to ensure estimation of non-singular/invertible class specific covariance matrices[2]. A popular rule-of-thumb is that there should be at least $5D$ data points/class for adequate estimation of the covariance matrix[23]. Existing hyperspectral classifiers including the BHC are thus susceptible to small sample size issues[32]. This paper introduces a technique for adaptively reducing the dimensionality of the feature space by recursively combining highly correlated, adjacent spectral bands until the reduced dimensionality is commensurate with the amount of data available. The more classic approach of regularization of covariance estimates is also embodied in this technique, which can be used in conjunction with both the BHC and error correcting output codes (ECOC). Experiments show that the resulting multi-classifier systems with adaptive feaure reduction modules provide substantial improvements over a range of small sample

sizes, without compromising performance for larger data sets.

2. Related Work

This paper addresses the issue of solving multi-class problems as well as tackling small sample size situations in the context of hyperspectral data. This section describes related work in these areas.

2.1. *Solving multi-class problems through output space decomposition*

Many real-life problems such as handwriting recognition involve more than two classes. In such cases, one has two choices: solve the problem directly using a single classifier that can provide multiple class labels, or decompose the output space into multiple two-class problems that are solved by different binary classifiers, and then combine the outputs of these classifiers in a suitable way to determine the final class label. A straightforward way of directly solving for multiple classes is to use a universal approximator such as the multi-layer perceptron (MLP) or radial basis function network, with C output units. A 1-of-C coding has to be used to obtain the desired outputs for training purposes, i.e., for a given input, the desired response is 1 for the output unit corresponding to its class label, and 0 for all the other output units. Theoretically, given such an encoding and a well trained network with an adequately large number of hidden units, it can be shown that the outputs of such a network approximate the corresponding aposteriori class probabilities[43,5] in the mean square sense. Thus, one can approach the Bayes optimum decision as closely as desired. In practice, with limited data, imperfect training and complex class boundaries, this becomes increasingly difficult to achieve as the number of classes increases.

Several other classification models such as decision trees (C5.0, CHAID, CART etc) can also directly address multi-class problems. However, several classifiers are more naturally suited to binary classification. A topical example is the support vector machine (SVM) in its original formulation[53]. Although several extensions of SVMs to multi-class problems have been subsequently suggested (see papers referred to in [19]), the results of [19] show that such direct approaches are inferior to decomposing the problem into several binary classification problems, each addressed by a binary SVM.

Over the years, several approaches to decomposing the output space, rather than directly solving for the C-class problem, have been proposed. These approaches can be categorized as:

1. solving C *one-versus-rest* two-class problems;
2. examining $\binom{C}{2}$ pairwise classifications,
3. applying error correcting output codes[12],
4. miscellaneous approaches, and

5. binary hierarchical classifiers.

We briefly summarize and compare the first four approaches before presenting the hierarchical output space decomposition approach proposed by us recently.

2.1.1. *One-versus-rest.*

The traditional approach to multi-class problems is to develop C classifiers, each focussed on distinguishing one particular class from the rest. Often this is achieved by developing a discriminant function for each of the C classes. A new data point is assigned the class label corresponding to the discriminant function that gives the highest value for that data point. For example, in Nilsson's classic linear machine[39], the discriminant functions are linear, so the decision boundaries are constrained to be hyperplanes that intersect at a point. This is an example of the *discriminant analysis* family of algorithms, that includes Quadratic Discriminant Analysis, Regularized Discriminant Analysis[13], and Kernel Discriminant Analysis[17]. The essential difference among different discriminant analysis methods is the nature and bias of the discriminant function used.

Anand et.al.[1] did a detailed empirical evaluation of the *one-versus-rest* method as compared to a one-shot approach when MLPs are used as classifiers. They showed that training an MLP with C output nodes, one for each class, was much slower than the modular alternative, at comparable levels of generalization performance. They also theoretically analyzed the spatial crosstalk phenomenon that hinders the one-shot approach. However, note that when C is large, there will be some classes for which the number of training data is much less than that for the rest of the classes combined, i.e. the corresponding two-class problem will encounter highly imbalanced priors. This leads to performance degradation and slower convergence, even for an MLP trained using error back-propagation, as shown for example, in [4].

2.1.2. *Pairwise classification.*

Also known as round robin classification[16], these approaches learn one classifier for each pair of classes (employing a total of $\binom{C}{2}$ classifiers in the process), and then combine the outputs of these classifiers in a variety of ways to determine the final class label. This approach has been investigated by several researchers[14,18,3,40]. Typically the binary classifiers are developed and in parallel, a notable exception being the efficient DAG structured ordering given in [40]. A straight-forward way of finding the winning class is through a simple voting scheme used for example in [14], which evaluates pairwise classification for two versions of CART and for the nearest neighbor rule. Alternatively, if the individual classifiers provide good estimates of the two-class posterior probabilities, then these estimates can be combined using an iterative hill-climbing approach suggested by [18].

Our first attempt at output space decomposition was to apply a pairwise classifier framework for land cover prediction problems involving hyperspectral data[28].

For this application, class-pair specific feature extraction not only yielded superior classification accuracies, but also provided important domain knowledge with regard to what features were more useful for discriminating specific pairs of classes. While such a modular learning approach for decomposing a C-class problem is attractive for a number of reasons, including focussed feature extraction, interpretability of results and automatic discovery of domain knowledge, the fact that it requires $\mathcal{O}(C^2)$ pairwise classifiers might make it less attractive for problems involving a large number of classes. Further, the combiner that integrates the results of all the $\binom{C}{2}$ classifiers must resolve the couplings among these outputs that might increase with the number of classes.

2.1.3. *Error correcting output codes (ECOC).*

Inspired by distributed output representations in biological systems, as well as by robust data communication ideas, ECOC is one of the most innovative and popular approaches to have emerged recently to deal with multi-class problems[12]. A C-class problem is encoded as $\bar{C}$ binary problems. For each binary problem, one subset of the classes serves as the positive class (target = 1) while the rest form the negative class (target = 0). As a consequence, each original class gets encoded into a $\bar{C}$ dimensional binary vector. The $C \times \bar{C}$ binary matrix is called the coding matrix. A given test input is labelled as belonging to the the class whose code is closest to the code formed by the outputs of the $\bar{C}$ classifiers in response to that input.

In the original ECOC paper[12], four ways of chosing the subsets of classes and therefore determining the code matrix were investigated. In general, it was believed that selecting matrices with good row and column separation would give better results. Most empirical studies using ECOCs employ a large number of binary classifiers ($\bar{C} \gg C$), and sometimes $\bar{C}$ is in the hundreds or more. Given such long codewords, subsets of classes chosen at random may perform nearly as well. In fact, the advantage of carefully crafted codewords seems to be clear only when the code lengths are short[57]. Moreover, the problem of designing an optimal binary code matrix is NP-Complete[10].

2.1.4. *Miscellaneous Approaches.*

There are some approaches to multi-class problems proposed by other authors that do not fall into the three categories described above. *Sequential methods* impose an ordering among the classes, and the classifiers are developed in sequence rather than in parallel. For example, one can first discriminate between class "1" and the rest. Then for data classified as "rest", a second classifier is designed to separate class "2" from the other remaining classes, and so on. Problem decomposition in the output space can also be accomplished implicitly by having C classifiers, each trying to solve the complete C-class problem, but with each classifier using input features most correlated with only one of the classes. This idea was used in [51] for creating

an ensemble of classifiers, each using different *input decimation*. This method not only reduces the correlation among individual classifiers in an ensemble, but also reduces the dimensionality of the input space for classification problems. Significant improvements in misclassification error, together with reduction in the number of features used, was obtained on various public domain datasets using this approach.

2.1.5. *The Binary Hierarchical Classifier.*

The top down BINARY HIERARCHICAL CLASSIFIER (TD-BHC) framework was introduced in [29,30] as a way of recursively decomposing a C-class problem into $C-1$ two-(meta)class problems. It results in a multi-classifier system with $C-1$ classifiers arranged as a binary tree. 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[29]. Fig. 1 shows an example of a C-class BHC. Note that the partitioning of a parent set of classes into two sets of metaclasses is not arbitrary, but is obtained through a deterministic annealing process that encourages similar classes to remain in the same partition. The tree structure also allows the easier discriminations to be accomplished earlier[21]. The TD-BHC was found to be competitive with pairwise classification and superior to a range of direct methods for classification of hyperspectral data[29]. Further results in [27] show that it performed well for several other data sets from UCI and NIST as well.

Subsequently, a bottom-up version (BU-BHC) was developed based on an agglomerative clustering algorithm used for recursively merging the two most similar meta-classes until only one meta-class remains[27]. Fisher's discriminant was again used as the distance measure for determining the order in which the classes are merged. The bottom-up procedure is computationally more expensive than the top-down version, but sometimes produces even better results, although it is locally more greedy.

Comments and Comparisons. A common characteristic of the first three approaches described above, is that they do not take into account the underlying affinities among the individual classes (for example, their closeness or amount of separation) while deciding on class selection/grouping for binary classification. Both one-versus-rest and pairwise methods treat each class the same way, while in ECOC, design of the code matrix is based on the properties of this matrix, rather than the classes they represent. That is why it is helpful to have a strong base learner when applying ECOC, since some of the groupings may lead to complicated decision boundaries. In contrast, the groupings in BHC are determined by the properties of the class distributions. Not being agnostic to class affinities helps us in determining natural groupings that facilitate both the discrimination process and the interpretation of results.

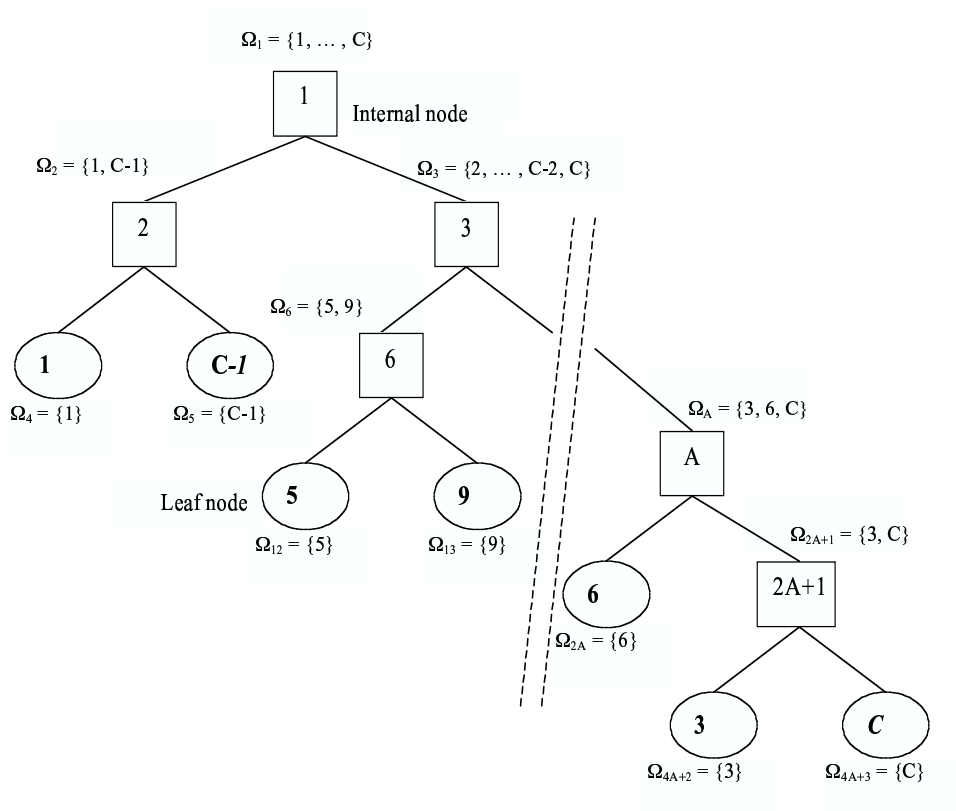

Fig. 1. An example of a BINARY HIERACHICAL (multi)-CLASSIFIER for solving a C-class problem. Each internal node n comprises of a feature extractor, a classifier, a left child $2n$, and a right child $2n+1$. Each node n is associated with a meta-class Ω_n.

Two noteworthy studies have emerged recently that compare one-versus-rest, pairwise and ECOC approaches. Furnkranz[16] shows that the $\binom{C}{2}$ learning problems of pairwise classification can be learned more efficiently than the C problems of the one-versus-rest technique. His analysis is independent of the base learning algorithm. He also observes that both of these approaches are more efficient than ECOC. A large number of empirical results are shown using Ripper and C5.0 as base classifiers. The BHC uses only $C-1$ classifiers, similar to one-versus-rest, but since the class groupings are based on affinities, the binary classifications are simpler in general. Hence BHCs do not compromise much on efficiency in the process of reducing the number of classifiers needed.

Hsu and Lin[19] completed a detailed study comparing one-versus-rest and pairwise classification, both using the SVM as base classifier, to two approaches for directly generalizing the SVM algorithm to multi-class problems. The pairwise method performed the best, both in terms of accuracy and training time. One-

versus-rest was second, and both methods were better than the direct generalizations of SVM.

2.2. *Small sample size problems*

The substantial methodology in this area can be largely categorized as one of four approaches[42]. Regularization methods, including shrinkage, try to stabilize the estimated covariance matrix directly by weighting the sample covariance matrix as well as supplemental matrices[47]. The covariance matrix can be shrunk toward the identity matrix or a pooled covariance matrix. Hybrid approaches assign weights to the sample covariance matrix and a pooled covariance matrix[48,36]. While this may reduce the variance of the parameter estimates, the bias of the estimates can increase dramatically. Rather than stabilizing the covariance matrix directly, the pseudo-inverse of the covariance matrix can be substituted for the true inverse. Pseudo-inversion utilizes the non-zero eigenvalues of the covariance matrix[15,47]. 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. Let $|X|$ represent the cardinality of the (training) set X. It has been shown that the pseudo-inverse performs best when $|X| = D/2$ and that the performance degrades as $|X|$ approaches D[46,41].

An alternate approach involves transforming the input space into a reduced feature space via feature extraction or selection[2,15]. Such transformations may result in some loss of interpretability and may be poorly estimated due to the limited data.

A third approach is to exploit an unlabelled examples that may be available using "semi-supervised learning" methods. Specific techniques for identifying and augmenting the existing training data with unlabeled data already exist and have been shown to enhance strictly supervised classification[49,20,6,24,37,11,44]. The quality of these approaches is very sensitive to the initial (guessed) labels of the unlabelled data, to the selection of the initial training samples and to outliers. Note that one can also artificially add labelled examples, sometimes called virtual examples, by perturbing the data or by exploiting any known invariances about the data[45].

The fourth approach uses an ensemble of weaker classifiers. Bagging, Simple Random Sub-sampling, Random Forests [8] and a variety of Arcing (Adaptively Reweighting and Combining techniques such as boosting) methods involve selecting subset samples of the original data and generating a classifier specific to each subsample[7]. When the data set is very small, however, these methods are inadequate because the degradation in individual classifier performance (because of lack of data) cannot be compensated for by the gains from using an ensemble[50].

2.3. *Feature Extraction from Hyperspectral Data*

Hyperspectral sensors simultaneously acquire information in hundreds of spectral bands. A hyperspectral image is essentially a three-dimensional array $I(p, q, d)$, where (p, q) denotes a pixel location in the image, and d denotes a spectral band

(wavelength). The value stored at $I(p, q, d)$ is the response (reflected or emitted energy) from the pixel (p, q) at a wavelength corresponding to spectral band d. The input space for a hyperspectral data (classification problem) is an ordered vector of real numbers of length D, the number of spectral bands, wherein the response of bands that are spectrally "near" each other tend to be highly correlated within certain regions of the spectrum.

Analysis of hundreds of simultaneous channels of data necessitates the use of either feature selection or extraction algorithms prior to classification. Feature selection algorithms for hyperspectral classification are costly, while feature extraction methods based on Karhunen Loeve (KL) transforms, Fisher's discriminant, or Bhattacharya distance cannot be used directly in the input space because the covariance matrices required by all these approaches are highly unreliable, given the ratio of the amount of training data to the number of input dimensions. The results are also difficult to analyze in terms of the physical characteristics of the individual classes and are not generalizable to other images.

Several authors have proposed approaches for extracting features from remotely sensed hyperspectral data[29,26,30,33]. Lee and Landgrebe[34,35] proposed methods for *feature extraction based on decision boundaries* for both Bayesian and neural network based classifiers. In these methods, a classifier is first learned for a two-class problem in the input space. A decision boundary is computed by moving along the closest samples in the two classes, and a vector normal to the decision boundary is noted. Eigenvectors of the decision boundary feature matrix formed by collection of these normal vectors yield the direction of projection for the two-class problem. The C-class problem is then solved using a (weighted) sum of the decision boundary feature matrices.

Jia and Richards proposed a Segmented Principal Components Transformation (SPCT) that exploits the observation that the original input features - the bands of the hyperspectral data - that are spectrally close to one another, tend to be highly correlated[25,26]. Edge detection algorithms are used to transform the original D individual bands into subsets of adjacent bands that are highly correlated, based on the estimated population correlation matrix. From each subset, the most significant principal components are selected to yield a feature vector that is significantly smaller in dimension than D. Although this approach exploits the highly correlated adjacent bands in hyperspectral data, it does not guarantee good discrimination capability because the Principal Component Transform preserves variance in the data rather than maximizing discrimination between classes. Additionally, the segmentation approach of SPCT is based on the correlation matrix over all of classes, and thus loses the often-significant variability in the class conditional correlation matrices. Subsequently, Kumar et al. proposed band combining techniques inspired by Best Basis functions[31]. 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[29]. Although these two methods utilize

the ordering of the bands and yield excellent discrimination, they are computationally intensive. Additionally, the quality of the discrimination functions, and thus the structure of the resulting feature space, is affected by the amount of training data, and this critical issue is not addressed.

3. An Adaptive Feature Space for Hyperspectral Data

We propose a simple method for tuning the amount of feature reduction to the quantity of available training data. The basic idea is to progressively merge adjacent bands that are highly correlated, so that the input dimensionality is reduced without significant loss in discriminatory power. While this method was originally designed for and is particularly suited to hyperspectral data[38], it can be applied to other high-dimensional data sets for which sequential inputs are highly correlated. We first describe how the technique is used in conjunction with the BHC, although the method can also be employed to reduce the input features used for other classifiers.

3.1. *Integrating Band Combination into Hierarchical, Multi-Classifier Systems*

The proposed approach can be viewed as a best-basis version of BHC (BB-BHC) that performs a band-combining step prior to the partitioning (top-down variant) or combining (bottom-up variant) of meta-classes. Band combining is performed 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 adjacent 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 pairwise correlations within that group:

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

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 whether 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. For linear models, the ratio of the number of training samples to the input dimensionality is

considered to be the most important indicator of whether the training set is adequate. Because the literature recommends different thresholds for the minimum $\alpha_{\text{ratio}} \leq \frac{|X|}{D}$, we allow this to be a user-defined input[32,22,42,54]. Note that $|X|$ represents the number of data points in a child meta-class, and this number decreases as we proceed toward 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 $l = 0$, $N = D$, and $B_l^k = [k, k]$, $\forall k = 1, \ldots, D$
3. If $N > D^*$ then continue. Otherwise, stop.
4. Find the best pair of band to merge: $K = \text{argmax}_{k=1,\ldots,N-1} Q(B_l^k \cup B_l^{k+1})$
5. Update band structure:
   - $l = l + 1$, $N = N - 1$
   - If $K > 1$ then $B_l^k = B_{l-1}^k$, $\forall k = 1, \ldots, K-1$
   - $B_l^K = B_{l-1}^K \cup B_{l-1}^{K+1}$
   - If $K < N$ then $B_l^k = B_{l-1}^{k+1}$, $\forall k = K+1, \ldots, N$
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 conditional covariance matrices, is dependent on the quantity of training data available to estimate the meta-class covariance matrices. This becomes even more critical for the low branches of the BB-BHC as the meta-classes become smaller in cardinality, and the amount of training data per meta-class decreases. 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 must be stabilized. The shrinkage technique[36] can be suitably employed for this purpose, taking advantage of the natural hierarchy provided by the BHC framework. We define the ancestor sample covariance matrix S^{Anc} as being the sample covariance matrix which is estimated from at least $\alpha_{\text{ratio}} D$ observations and is most closely related to L_k based on the BB-BHC structure. Because the trees can be constructed either in a top-down or bottom-up manner, the search for S^{Anc} must be performed differently for the two approaches. 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 $|X_{\Omega_k}| \leq \alpha_{\text{ratio}} D$, 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 now being constructed bottom-up, the structure cannot be climbed in search of a suitable S^{Anc}. Therefore, if $|X_{\Omega_i+\Omega_j}| \leq \alpha_{\text{ratio}} D$, then $S^{\text{Anc}} = \sum_{i=1}^{C} P(L_i)S^{L_i}$. Note that this estimate for S^{Anc} 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 utilized to estimate the correlation measure in (1).

4. Empirical Studies

4.1. *Multiple Classifier Systems Studied.*

The base methods for comparison are the bottom-up and top-down versions of BHC, denoted by TD-BHC and BU-BHC respectively. Applying the *pseudo-inverse* for tree construction (estimating Fisher's discriminant) and feature extraction (calculating Fisher's linear discriminant function) yields TD-P-BHC, BU-P-BHC, while using the adaptive best-basis construction results in TD-BB-BHC and BU-BB-BHC. We also wanted to compare these methods with other approaches to multi-class problems. Previously we had shown that, for a hyperspectral data set with at least 180 samples per class, the TD-BHC gives comparable results to a pairwise classifier architecture that utilizes a best-basis technique for combining bands[29]. For smaller sample sizes, the pairwise architecture is expected to suffer even more that BHC or ECOC because each of its $\binom{C}{2}$ component binary classifiers can only use data from two of the original classes. In contrast, the BHC deals with meta-classes at all levels above the leaf classifiers. At the root, all the data are available, and the amount of data available at each internal node progressively diminishes, as the number of original classes in each metaclass decreases. Thus, it is clear that the pairwise or round-robin architecture will be at a further disadvantage as the sample sizes shrink, and hence is not considered in this study.

Instead we compare the results with those obtained by an ECOC architecture. As mentioned earlier, each component classifier in this framework solves a binary problem, with the original set of classes divided into two groups based on the corresponding column of the code matrix. Thus each classifier uses all the data! Moreover, if the code matrix is chosen so that the two groups always have roughly the same number of classes, and the priors of these classes are comparable, then each two-metaclass problem is reasonably balanced. Thus one would expect the ECOC method to be much less susceptible to small size problems as compared to pairwise classification.

We decided to use the Bose-Chaudhuri-Hochquenghem (BCH) code, which shows excellent separation among both rows and columns. BCH codes are multilevel, cyclic, error-correcting, variable-length digital codes used to correct errors up to approximately 25% of the total number of digits. They are based on Galois field theory and have superior error correcting properties to well-known Hamming

Table 1. BCH Table

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	0	1	1	0	0	1	0	1	0	0	0	0	1
3	0	1	1	0	1	0	1	1	1	1	0	0	0	1	0
4	1	0	1	1	0	0	1	0	1	0	0	0	0	1	1
5	1	1	0	1	0	1	1	1	1	0	0	0	1	0	0
6	0	0	0	0	1	1	1	0	1	1	0	0	1	0	1
7	1	0	1	1	1	1	0	0	0	1	0	0	1	1	0
8	0	1	1	0	0	1	0	1	0	0	0	0	1	1	1
9	0	1	1	1	0	1	1	0	0	1	0	1	0	0	0
10	1	0	1	0	1	1	1	1	0	0	0	1	0	0	1
11	0	0	0	1	1	1	0	1	1	0	0	1	0	1	0
12	1	1	0	0	0	1	0	0	1	1	0	1	0	1	1
13	1	0	1	0	0	0	0	1	1	1	0	1	1	0	0
14	0	1	1	1	1	0	0	0	1	0	0	1	1	0	1
15	1	1	0	0	1	0	1	0	0	0	0	1	1	1	0
16	0	0	0	1	0	0	1	1	0	1	0	1	1	1	1

codes, and were also recommended in the original ECOC paper[12] for a larger number of classes. The two datasets studied in this paper involve 11 and 13 classes respectively. We chose to use a code-length of 15 to accommodate up to 32 classes. This choice provides an error correction of 3 bits. Therefore any two rows have a Hamming distance of at least 7 bits. The first 16 rows of the BCH code for this choice are shown in Table 1. For the experiments, the first 11 and first 13 rows were chosen respectively. Note that the most significant data bit is all zeros (it is all ones for the next 16 entries), and thus this column (number 11 in Table 1) is deleted, leaving 14 binary classifiers. This number is comparable to the $C-1$ classifiers used in the BHC.

4.2. *Empirical Results*

The proposed algorithms were tested on hyperspectral data obtained from two sites: Bolivar Peninsula, Galveston, Texas and NASA's John F. Kennedy Space Center (KSC), Florida.

4.2.1. *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. For classification purposes, 11 classes representing the various land cover types were defined for the site (Table 2). These include: water, wetlands (low proximal marsh, high proximal marsh, high distal marsh, and pure *salicornia*) and uplands (trees, general uplands, two agricultural classes, sand flats, and a transition zone)[52,56]. The low proximal marsh corresponds to tidal flats

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

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

comprised of *Spartina alterniflora*, which experiences frequent flooding. The high proximal marsh, which is composed of a mixture of *Spartina alterniflora* and *Salicornia virginica*, is flooded less frequently and has more continuous vegetation cover. The high distal marsh, which is inundated even less frequently than the proximal marshes, contains *Spartina patens, Salicornia virginica* and *Juncus roemerianus.* Adjacent to the high distal marsh, a small highly saline region of sand flats surrounded by pure *Salicornia virginica* delineates the boundary between the wetlands and uplands. The topography of these areas is mainly a function of sedimentary processes such as high-energy wave and low-energy tidal and wind processes. As a result, the frequency of the inundation, soil salinity, and vegetation cover all depend on this topography[56]. HyMap (Hyperspectral Mapper) collected data over Bolivar Peninsula on September 17, 1999, at 5m spatial resolution. Data were acquired in 126 bands with almost contiguous spectral coverage from 440-2480 nm[9]. After removing water absorption and low SNR bands, 122 bands were used in the analysis.

Multiple experiments were performed using stratified (class specific) sampling at percentages of: 75, 50, 30, 15, and 5. Even at the sampling percentage of 75, the amounts of training data for classes 5 and 10 are still less then D (sand flats $|X_{L_5}| = 111$ and transition zone $|X_{L_{10}}| = 86$). We used $\alpha_{\text{ratio}} = 5$ for all sampling percentages. Ten experiments, using simple random sampling of the training data, 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 Fishers discriminant as a distance measure) and feature extraction (calculating Fishers linear discriminant function) [TD-P-BHC, BU-P-BHC], and the adaptive best-basis BHC [TD-BB-BHC, BU-BB-BHC]. Ten additional experiments were conducted using a best basis implementation of ECOC [BB-ECOC] for comparison. The results are presented in Figure 2. Each data point in Figure 2 (top) denotes the mean value of test set accuracy. The corresponding standard devations are shown separately in Figure 2 (bottom) to

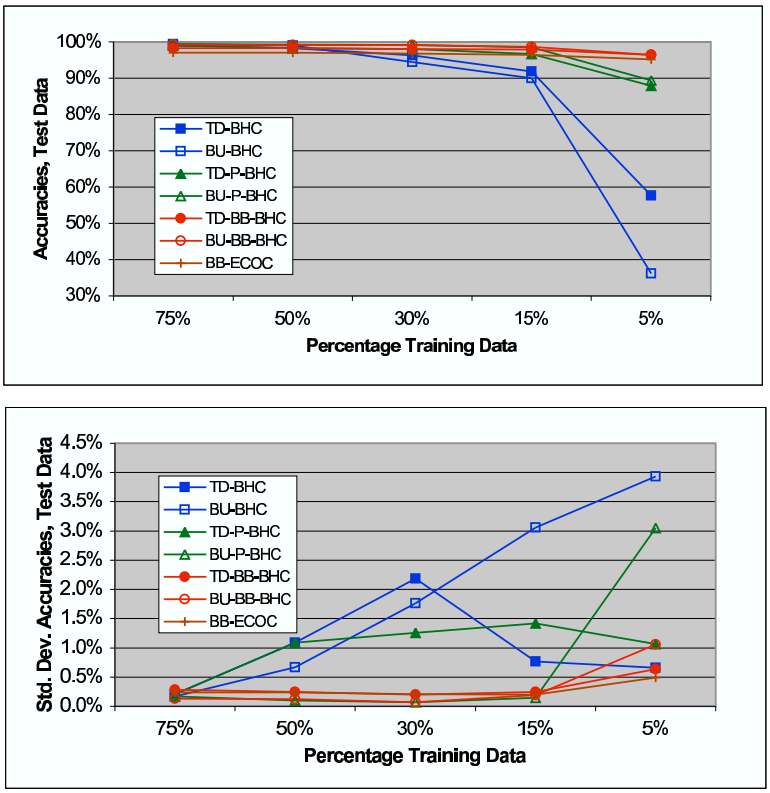

Fig. 2. Classification (test set) accuracies for Bolivar Peninsula

reduce clutter. Note that for getting more reliable accuracy estimates, all the data not in a given training sample was used for testing, in both sets of experiments. This does introduce an artifact however that for smaller sampling rates, there is more test data and hence the standard deviations are lower than what one expects from smaller test sets. Also to have a fair comparison, for each run, the same specific training/test split was used for the first six models. However, the experiments on ECOC were conducted later, when these splits were not available, so new random splits had to be obtained. Since the results are averaged over ten runs, we do not expect that this has had any significant affect on the comparisons.

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 an extremely low number of observations. At 75% sampling, the performance of all 7 classifiers is comparable in terms of both the average and standard deviations of the accuracies of the test data. At 50% sampling, which is typically used to separate data sets into training and testing, the average overall accuracies of the classifiers are still similar. The variability of the BHC, increases somewhat, as does the TD-P-BHC. Importantly, even though using the pseudo-inverse does not improve the average

accuracies at 50% sampling, because there are at least $D + 1$ observations per L_i, the results indicate that while the covariance matrices are non-singular, they are still poorly estimated. Not only does the BB-BHC perform the best at every sampling percentage with respect to the other TD and BU classifiers, but the accuracies are generally more stable (smaller standard deviation of accuracies) as well. The BB-BHC also yields slightly higher accuracies than the BB-ECOC, even with the classifier diversity introduced by the ECOC. However, as expected, the ECOC produces extremely stable results, as indicated by the standard deviation of the accuracies at each sampling percentage. Thus, combating the limited amount of training data by using the correlation matrix for feature reduction helps retain the information necessary for successful land cover prediction in both the structured BHC and the ECOC. Overall classification accuracies of $> 90\%$ can still be achieved at the 5% sampling rate.

4.2.2. *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. The impoundments were created during the 1950s and 1960s for the purpose of mosquito control. The marshes along the IRL contain both high and low marsh communities. The three dominant marsh groups that comprise the high marsh communities are cabbage palm savanna, sand cordgrass, and black rush. The cabbage palm savanna consists of isolated canopies of Cabbage Palm (*Sabal palmetto*) and a graminoid layer of sand cordgrass (*Spartina bakerii*) and black rush marsh (*Juncus roemerianus*). Salt tolerant grasses and halophytes dominate the low marsh communities. The primary salt tolerant grass is *Distichilis spicata*. Halophytes typically include *Batis maritima* and *Salicornia virginica*. This study also includes investigation of upland vegetation, as it is adjacent to the impounded wetlands. In addition, accurate classification and mapping of upland vegetation is important for monitoring habitat of the endangered Florida Scrub Jay. The majority of the upland vegetation at KSC is oak scrub and saw palmetto scrub. Other upland communities include slash pine (*Pinus elliottii*) and hardwood swamps that are dominated by deciduous trees such as Red Maple (*Acer rubrum*). Dense hammocks of Cabbage Palm (*S. palmetto*) and Live Oaks (*Quercus virginiana*) are also common[55]. Discrimination 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 3).

The NASA AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) spectrometer acquired data over the KSC, Florida on March 23, 1996. AVIRIS acquires

Table 3. Classes for Kennedy Space Center and the quantity of training data per class

Class	Name	Total Observation
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

data in 224 bands of 10 nm width from 400 - 2500 nm. The KSC data, acquired from an altitude of approximately 20 km, have a spatial resolution of 18 m. After removing water absorption and low SNR bands, $D = 176$ bands were used for the analysis. Again, multiple experiments were performed using stratified (class specific) sampling at percentages of: 75, 50, 30, 15, and 5. At 75% sampling rate, the quantity of training data for classes 5, 6, and 7 is less than D and, at 50%, so are classes 2, 3, and 4. Ten experiments, using simple random sampling, were performed at each percentage for all seven classifiers. The results are presented in Figure 3.

The overall trends in test set accuracies for Cape Canaveral are very similar to those of Bolivar Peninsula, although the performance of the BHC degrades even more quickly. At the lower sampling percentages, the covariance matrices of the BHC 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 increase dramatically at the 15%-30% sampling rate for the pseudo-inverse classifiers where the covariance matrices are still helping maintain a higher level of classification accuracy (than the 5% range), though unstable. Although average accuracies are somewhat lower than those produced by the BB-BHC, the BB-ECOC yields quite stable results at all sampling levels.

5. Conclusions and Future Work

The dependency of classification accuracy on the ratio of training data size to the dimensionality of the data has been widely noted and needs to be addressed in the design of a classifier. While the advent of hyperspectral sensors has provided unique opportunities in remote sensing, the high-dimensional features provided by these sensors signify that researchers should take note of classifiers that are designed to

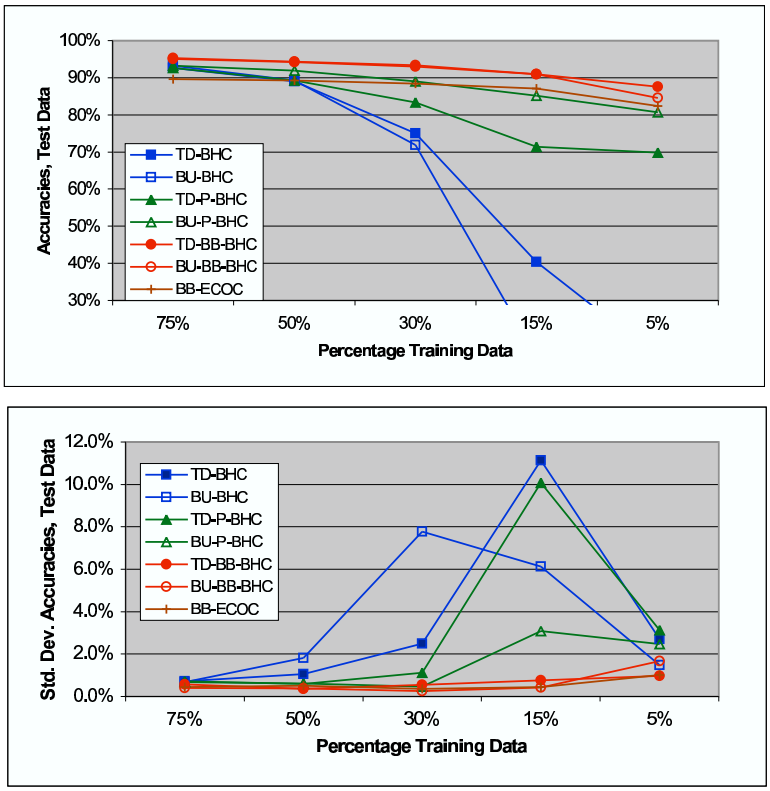

Fig. 3. Classification (test set) accuracies for Cape Canaveral

be more tolerant of the quantity of training data available. This paper presented a multiple classifier framework that utilizes the flexibility gained by transforming the output space and input space simultaneously to combat both the small sample size problem and the issue of being faced with a large number of classes. By reducing the size of the feature space in a directed manner, dependent on the quantity of training data available in the binary hierarchy of meta-classes, a high level of classification accuracy is preserved even when the quantity of training data for some classes is low. In addition, the adaptive feature space technique can be used with other multiple classifier approaches to multi-class problems, most notably, the error correcting output code technique.

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 bands importance, the interpretation and insight should be improved if feature selection

were performed rather than feature extraction. Therefore, the use of feature selection rather than feature extraction, and the ensuing trade-off between classification accuracy and retention of domain knowledge, should be investigated further.

Another contribution of this paper is that it provides an overview of several approaches to multiple-class problems, and also provides the first comparison of the BHC method with the powerful and popular ECOC approach. The results are quite flattering to the BHC, at least for the two challenging hyperspectral datasets that we examined.

Most likely, this is due to the grouping of classes based on affinities rather than on a code matrix that does not consider the properties of the individual classes. However, this advantage is perhaps amplified because the base classifiers used in this study are not very powerful. One needs to note that the design space is indeed very broad for both methodologies. For example, the ECOC can be used with a variety of base classifiers and feature selection/extraction methods, and there are several ways of obtaining suitable coding matrices as well[57]. Similarly other types of classifiers and feature extraction modules can be organized in a hierarchical fashion as well. A very large number of experiments need to be performed to explore this rich design space. Finally, while the focus of this paper was on hyperspectral data classification, one also needs to experiment with other types of data sets, such as letter recognition, that exhibit a moderately large number of classes and fairly high input dimensionality, to fully flush out the scope and power of the methodologies proposed in this paper.

Acknowledgments

This research was supported in part by the NASA EO -1 program, Grant NCC5-463, the Terrestrial Sciences Program of the Army Research Office(DAAG55-98-1-0287), NSF grant ECS-9900353, the Texas Advanced Technology Research Program(CSRA-ATP-009), and grant No. 8032 from Intel Corp. We thank Amy Neuenschander and Yangchi Chen for their help with data preparation and interpretation of results.

References

1. R. Anand, K. Methrotra, C. K. Mohan, and S. Ranka. Efficient classification for multiclass problems using modular neural networks. *IEEE Transactions on Neural Networks*, 6(1):117–125, 1995.
2. T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, New York, 1984.
3. C. Burges B. Schlkopf and A. Smola, editors. *Advances in Kernel Methods - Support Vector Machines*. MIT Press, 1998.
4. E. Barnard and E.C. Botha. Back-propagation uses prior information efficiently. *IEEE Transactions on Neural Networks*, 4(5):794–802, 1993.
5. C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, New York, 1995.

6. A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings 11th Annual Conference on Computational Learning theory*, pages 92–100, 1998.
7. L. Breiman. Combining predictors. In A. Sharkey, editor, *Combining Artificial Neural Nets*, pages 31–50. Springer-Verlag, 1999.
8. L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
9. T. Cocks, R. Jenssen, A. Stewart, I. Wilson, and T. Shields. The HYMAP airborne hyperspectral sensor: the system, calibration and performance. In *Proc. 1st EARSeL Workshop on Imaging Spectroscopy (M. Schaepman, D. Schlpfer, and K.I. Itten, Eds.), Zurich, EARSeL, Paris*, pages 37–42, October 1998.
10. K. Crammer and Y. Singer. On the learnability and design of output codes for multiclass problems. In *Computational Learning Theory*, pages 35–46, 2000.
11. V. R. De Sa. Learning classification with unlabeled data. In J. D. Cowan, G. Tesauro, and J. Alspector, editors, *Neural Information Processing Systems*, San Francisco, CA, 1994. Morgan Kaufmann.
12. T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2(1):263–286, 1995.
13. J. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.
14. J. Friedman. Another approach to polychotomous classification. Technical report, Stanford University, 1996.
15. K. Fukunaga. *Introduction to Statistical Pattern Recognition (2nd Ed.)*. Academic Press, 1990.
16. J. Furnkranz. Round robin classification. *Journal of Machine Learning Research*, 2:721–747, 2002.
17. J. Ghosh and S.V. Chakravarthy. The rapid kernel classifier: A link between the self-organizing feature map and the radial basis function network. *Journal of Intelligent Material Systems and Structures*, 5:211–219, 2 1994.
18. T. Hastie and R. Tibshirani. Classification by pairwise coupling. In *Proceedings of the International Joint Conference on Neural Networks*, Washington, D.C., July 1999.
19. C. W. Hsu and C. J. Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.
20. Q. Jackson and David Landgrebe. An adaptive classifier design for high-dimensional data analysis with a limited training data set. *IEEE Transactions on Geoscience and Remote Sensing*, 39(12):2664–2679, 2001.
21. A. K. Jain. Advances in statistical pattern recognition. In F. A. Devijver and J. Kittler, editors, *Pattern Recognition Theory and Applications*, pages 1–19. Springer-Verlag, 1986.
22. A.K. Jain, P.W. Duin, and J. Mao. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.
23. A.K. Jain, R.P.W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Trans. PAMI*, 22(1):4–37, Jan 2000.
24. B. Jeon and D. Landgrebe. partially supervised classification using weighted unsupervised clustering. *IEEE Transactions on Geoscience and Remote Sensing*, 37(2):1073–1079, 1999.
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 Transactions on*

Geoscience and Remote Sensing, 37(1):538–542, January 1999.

27. S. Kumar. *Modular learning through output space decomposition*. PhD thesis, Dept. of ECE, Univ. of Texas at Austin, Dec., 2000.
28. S. Kumar, M. M. Crawford, and J. Ghosh. A versatile framework for labelling imagery with a large number of classes. In *Proc. IJCNN*, 1999.
29. S. Kumar, J. Ghosh, and M. M. Crawford. Hierarchical fusion of multiple classifiers for hyperspectral data analysis. *Pattern Analysis and Applications, spl. Issue on Fusion of Multiple Classifiers*, 5(2):210–220, 2002.
30. S. Kumar, J. Ghosh, and M.M. Crawford. A hierarchical multiclassifier system for hyperspectral data analysis. In J. Kittler and F. Roli, editors, *Multiple Classifier Systems*, pages 270–279. LNCS vol 1857, Springer, 2000.
31. S. Kumar, J. Ghosh, and M.M. Crawford. Best basis feature extraction algorithms for classification of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 39(7):1368–1379, July 2001.
32. D. A. Landgrebe. Information extraction principles and methods for multispectral and hyperspectral image data. In C. H. Chen, editor, *Information Processing for Remote Sensing*, pages 3–37. World Scientific Publishing Co., Inc., 1999.
33. D.A. Landgrebe. Hyperspectral image data analysis as a high dimensional signal processing problem. *Special Issue of the IEEE Signal Processing Magazine*, 19(1):17–28, 2002.
34. C. Lee and D. A. Landgrebe. Decision boundary feature extraction for neural networks. *IEEE Transactions on Neural Networks*, 8(1):75–83, January 1997.
35. C. Lee and D.A. Landgrebe. Feature extraction based on decision boundaries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(4):388–400, April 1993.
36. A. McCallum, R. Rosenfeld, T. Mitchell, and A.Y. Nigam. Improving text classification by shrinkage in a hierarchy of classes. In *Proc. 15th International Conf. on Machine Learning*, pages 359–367. Morgan Kaufmann, San Francisco, CA, 1998.
37. T.M. Mitchell. The role of unlabeled data in supervised learning. In *Proceedings 6th International Colloquium on Cognitive Science*, 1999.
38. 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 J. Kittler and F. Roli, editors, *Multiple Classifier Systems*, pages 189–200. LNCS vol 2364, Springer, 2002.
39. N. J. Nilsson. *Learning Machines: Foundations of Trainable Pattern-Classifying Systems*. McGraw Hill, NY, 1965.
40. J. C. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin DAGs for multiclass classification. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Neural Information Processing Systems*, volume 12, pages 547–553. MIT Press, 2000.
41. S.J. Raudys and R.P.W. Duin. Expected classification error of the fisher linear classifier with pseudo-inverse covariance matrix. *Pattern Recognition Letters*, 19:385–392, 1998.
42. S.J. Raudys and A.K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3):252–264, 1991.
43. M.D. Richard and R.P. Lippmann. Neural network classifiers estimate Bayesian a posteriori probabilities. *Neural Computation*, 3(4):461–483, 1991.
44. 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 Transactions on Geoscience and Remote Sensing*, 32(5):1087–95, 1994.

45. Patrice Simard, Yann LeCun, John S. Denker, and Bernard Victorri. Transformation invariance in pattern recognition-tangent distance and tangent propagation. In *Neural Networks: Tricks of the Trade*, pages 239–27, 1996.
46. M. Skurichina. *Stabilizing weak classifiers*. PhD thesis, Vilnius State Univesity, 2001.
47. M. Skurichina and R.P.W. Duin. Stabilizing classifiers for very small sample sizes. In *Proceedings 13th International Conference on Pattern Recognition (Vienna, Austria, Aug.25-29) Vol. 2, Track B: Pattern Recognition and Signal Analysis, IEEE Computer Society Press, Los Alamitos,*, pages 891–896, 1996.
48. S. Tadjudin and D.A. Landgrebe. Covariance estimation with limited training samples. *IEEE Transactions on Geoscience and Remote Sensing*, 37(4):2113–2118, 1999.
49. S. Taudjudin and D.A. Landgrebe. Robust parameter estimation for mixture model. *IEEE Transactions on Geoscience and Remote Sensing*, 38(1):439–445, 2000.
50. K. Tumer and J. Ghosh. Error correlation and error reduction in ensemble classifiers. *Connection Science*, 8(3-4):385–404, December 1996.
51. K. Tumer and N. C. Oza. Decimated input ensembles for improved generalization. In *Proceedings of the International Joint Conference on Neural Networks*, Washington, D.C., 1999.
52. R. Ustun. Spectral/spatial classification and output-based fusion for multisensor remotely sensed imaged data. Master's thesis, The University of Texas at Austin, Austin, TX, December 2000.
53. Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.
54. Andrew Webb. *Statistical pattern recognition*. Oxford University Press, London, 1999.
55. Webpage. Center for Space Research, University of Texas at Austin http://www.csr.utexas.edu/rs/research/ksc/index.html.
56. W.A. White, T.R. Calnan, R.A. Morton, R.S. Kimble, T.G. Littleton, J.H. McGowen, H.S. Nance, and K.E. Schmedes. *Submerged Lands of Texas, Galveston-Houston Area: Sediments, Geochemistry, Benthic Microinvertebrates, and Associated Wetlands*. Bureau of Economic Geology, University of Texas at Austin, 1985.
57. T. Windeatt and R. Ghaderi. Binary labelling and decision level fusion. *Information Fusion*, 2(2):103–112, 2001.