Methodology for Hyperspectral Band and Classification Model Selection

Peter Groves
National Center for Supercomputing Applications
University of Illinois Urbana-Champaign
Champaign, Illinois, 61820
Email: pgroves@uiuc.edu

Peter Bajcsy
National Center for Supercomputing Applications
University of Illinois Urbana-Champaign
Champaign, Illinois, 61820
Email: pbajcsy@ncsa.uiuc.edu

***Abstract*— Feature selection is one of the fundamental problems in nearly every application of statistical modeling, and hyperspectral data analysis is no exception. We propose a new methodology for combining unsupervised and supervised methods under classification accuracy and computational requirement constraints. It is designed to perform not only hyperspectral band (wavelength range) selection but also classification method selection. The procedure involves ranking bands based on information content and redundancy and evaluating a varying number of the top ranked bands. We term this technique Rank Ordered With Accuracy Selection (ROWAS). It provides a good tradeoff between feature space exploration and computational efficiency. To verify our methodology, we conducted experiments with a georeferenced hyperspectral image (acquired by an AVIRIS sensor) and categorical ground measurements.**

I. Introduction

For tasks involving classification of remotely sensed imagery, hyperspectral sensors can provide a wealth of useful information. Unfortunately, the efficacy of standard classification techniques applied to such data often is hindered by redundant or irrelevant information present in the data set [1],[2]. Furthermore, when the algorithms scale poorly to the large number of dimensions inherent in hyperspectral imagery, prohibitive computational requirements can emerge.

In most application areas, the goal of hyperspectral image analysis is to classify or discriminate objects. Driven by classification or discrimination accuracy, one would expect that, as the number of hyperspectral bands increases, the accuracy of classification should also increase. Nonetheless, this is not the case in a model-based analysis [1], [3]. Redundancy in data can cause convergence instability of models. Furthermore, spectral value variations due to noise in redundant data propagate through a classification or discrimination model. The same is true of spectral information that has no relation to the feature being classified in the underlying mathematical model. Such information is the same as noise to any statistical model, even if it is novel witin the data set and accurate. Thus, processing a large number of hyperspectral bands can result in higher classification error than processing a subset of relevant bands without redundancy. In addition, computational requirements for processing large hyperspectral data sets might be prohibitive when using modeling techniques that scale poorly with the number of features. A method for selecting a data subset is therefore sought.

The driving force of our algorithm is the desire to find the combination of bands and classification model that minimizes an error function. Our first consideration is the No Free Lunch Theorem [4], which states that over all sets of problems, all classification models have the same accuracy; classification models can only have better performance on a case by case basis. If we make no assumption of the underlying relationship between the hyperspectral data and the predicted variable(s), it follows that we must try an assortment of modeling techniques if we hope to achieve high accuracy.

If we wished to guarantee finding the absolute optimal accuracy, we would exhaustively try all possible combinations of bands, in all available classification algorithms, and pick the best. This is infeasible, of course, as the number of possible subsets of bands is $2^n - 1$, where n is the number of bands, and there are potentially hundreds or even thousands of bands in hyperspectral imagery. In another possible approach, one could attempt to use a standard optimizer (hill climbing, genetic algorithm, etc.) to find the optimal set of bands by minimizing the classification error a set of bands produces with a given classifier. When we consider the time it takes to do a single evaluation of a set of bands, and the fact that it often takes many thousands of evaluations for an optimizer to reach a good result (if random restarting is necessary to break out of local optima), we find that the computational requirements are still prohibitively high for any non-trivial data set.

In this work, we propose a method to overcome these obstacles by ranking bands based on their information content and distinctiveness. A wrapper method, where subsets of features are evaluated based on cross-validated accuracy scores [5], is then used to determine the optimal number of top ranked bands to use.

II. Methodology

To overcome the computational and combinatorial restrictions, we propose to rank order the hyperspectral bands in terms of information content, and then evaluate the classification error of the i top bands at a time, where i starts near 1 and is incremented by some small value (in our experiments, i starts at 2 and has an increment size of 2). Using this method of evaluating top ranked bands, we can reformulate our main

optimization problem to that of finding the best combination of ranking method, classification model, and number of top bands to use (with the wavelengths of the top bands determined by the ranking method). This reduces the problem to one that is on the cusp of being computationally infeasible, taking an amount of time on the order of a few days when run on multiple modern computers for a 200 band data set. For convenience, we refer to the general technique of ranking features, then adding a few features at a time to a set to be evaluated by cross-validation, as Rank Ordered With Accuracy Selection, or ROWAS.

In this work, seven unsupervised ranking methods and three supervised classification methods are explored. The ranking methods can loosely be grouped into two categories. The first are those based on measures of a individual band's information content. The most straightforward is a common entropy measure. The other technique that falls into this category is our *spatial contrast* measure, which indicates the level of discrimination a band provides if we consider every pair of spatially adjacent data points to belong to some differing, unknown categories. The other category consists of those methods based on redundancy among multiple bands. These methods work mainly by penalizing bands for being similar to others, and then selecting those that are least penalized. Included in this category are methods based on the correlation between pairs of bands, the predictability of one band based on the bands adjacent to it in the spectrum, a band's contribution to a principal components analysis, and the degree to which a pair of bands' spectral ratio differs from the average spectral ratio over all pairs of bands. The supervised methods are naïve bayes, C4.5 decision tree, and k-nearest neighbors. These are common classification methods and are described in many machine learning and pattern recognition texts such as [4].

The following two sections present detailed descriptions of both the unsupervised (III) and supervised (IV) methods.

III. Unsupervised Methods

A. Information Entropy

This method is based on evaluating each band separately using the information entropy measure ([6], chapter 3) defined below.

$$H(\lambda_i) = -\sum_{k=1}^{m} p(\phi_k^i) \ln p(\phi_k^i) \tag{1}$$

$$p(\phi_k^i) = p(\min_k^i \leq I(\lambda_i) < \max_k^i) \tag{2}$$

H is the entropy measure of wavelength λ_i. Eq. 2 merely formalizes that the probability distribution function of the intensity value $I(\lambda_i)$ is estimated via a histogram where each bin k used to estimate a probability for a range of values of $I(\lambda_i)$ is defined by $\{\min_k^i, \max_k^i\}$. m is the number of bins used in each histogram. Generally, if the entropy value H is high then the amount of information in the data is large. Thus, the bands are ranked in the ascending order from the band with the highest entropy value (large amount of information) to the band with the smallest entropy value (small amount of information).

B. First Spectral Derivative

The bandwidth, or wavelength range, of each band is a variable in a hyperspectral sensor design [2], [7]. This method explores the bandwidth variable as a function of added information. It is apparent that if two adjacent bands do not differ greatly then the underlying geo-spatial property can be characterized with only one band. The mathematical description is shown below in (3), where I represents the hyperspectral reflectance value of central wavelength λ_i at spatial location x. Thus, if D_1 is equal to zero then one of the bands is redundant. In general, the adjacent bands that differ significantly should be retained, while similar adjacent bands can be reduced.

$$D_1(\lambda_i) = \sum_x |I(x, \lambda_i) - I(x, \lambda_{i+1})| \tag{3}$$

C. Second Spectral Derivative

Similar to the first spectral derivative, this method explores the bandwidth variable in hyperspectral imagery as a function of added information. If three bands are adjacent, and the outer bands can be used to predict the center band through linear interpolation, then the center band is redundant. The larger the deviation from a linear model, the higher the information value of the band. The mathematical description of this method is shown below, where D_2 represents the measure of linear deviation of a central wavelength λ_i when I is the reflectance value of central wavelength λ_j at spatial location x for the appropriate values of j.

$$D_2(\lambda_i) = \sum_x |I(x, \lambda_{i-1}) - 2I(x, \lambda_i) + I(x, \lambda_{i+1})| \tag{4}$$

D. Contrast Measure

This method is based on the assumption that each band could be used for classification purposes by itself. The usefulness of a band would be measured by a classification error achieved by using only the band under consideration and minimizing the error. In order to minimize a classification error, it is desirable to select bands that provide the highest amplitude discrimination (image contrast) among classes. If the class boundaries were known *a priori* then the measure would be computed as a sum of all contrast values along the boundaries. However, the class boundaries are unknown *a priori* in the unsupervised case. One can evaluate contrast at all spatial locations instead assuming that each class is defined as a homogeneous region (no texture variation within a class). The mathematical description of the contrast measure computation is shown below for a discrete case.

$$ContrastM(\lambda) = \sum_{i=1}^{m} |f_i - E(f)| * f_i \tag{5}$$

f is the histogram (estimated probability density function) of all contrast values computed across one band by using a Sobel edge detector ([6], Chapter 4). $E(f)$ is the sample mean of the histogram f and is the central wavelength. m is the number of distinct contrast values in a discrete case. The

equation includes the contrast magnitude term and the term with the likelihood of contrast occurrence. In general, bands characterized by a large value of $ContrastM$ are ranked higher (good class discrimination) than the bands with a small value of $ContrastM$.

E. Spectral Ratio Measure

In many practical cases, band ratios are effective in revealing information about inverse relationship between spectral responses to the same phenomenon (e.g., living vegetation using the normalized difference vegetation index ([8], Chapters 16.6 and 17.7). This method explores the band ratio quotients for ranking bands and identifies bands that differ just by a scaling factor. The larger the deviation from the average of ratios $E(ratio)$ over the entire image, the higher the $RatioM$ value of the band. The mathematical description of this method is shown below, where $RatioM$ represents the measure and I is the reflectance value of central wavelength λ_j at spatial location x.

$$RatioM(\lambda_i) = \sum_x \left| \frac{I(x,\lambda_i)}{I(x,\lambda_{i+1})} - E\left(\frac{I(x,\lambda_i)}{I(x,\lambda_{i+1})}\right) \right| \quad (6)$$

F. Correlation Measure

One of the standard measures of band similarity is normalized correlation [4]. The normalized correlation metric is a statistical measure that performs well if a signal-to-noise ratio is large enough. The correlation based band ordering computes the normalized correlation measure for all pairs of bands similar to the spatial autocorrelation method applied to all ratios of pairs of image bands in [9]. Considering all pairs of bands and not just those that are spatially adjacent is an important distinction of the correlation based method. The mathematical description of the normalized correlation measure is shown below, where $CorM(\lambda_i, \lambda_j)$ represents the measure and I is the reflectance value of central wavelength λ. E denotes an expected value and σ is a standard deviation.

$$CorM(\lambda_i, \lambda_j) = \frac{E(I(\lambda_i) * I(\lambda_j)) - E(I(\lambda_i)) * E(I(\lambda_j))}{\sigma(I(\lambda_i)) * \sigma(I(\lambda_j))} \quad (7)$$

After selecting the first least correlated band based on all other bands, the subsequent bands are chosen as the least correlated bands with the previously selected bands. This type of ranking is based on mathematical analysis of [10], where spectrally adjacent blocks of correlated bands are represented in a selected subset.

G. Principal Component Analysis Ranking (PCAr)

Principal component analysis has been used very frequently for band selection in the past [8]. The method transforms a multidimensional space to one of an equivalent number of dimensions where the first dimension contains the most variability in the data, the second the second most, and so on. The process of creating this space gives two sets of outputs. The first is a set of values that indicate the amount of variability each of the new dimensions in the new space represents, which are also known as eigenvalues (ϵ). The second is a set of vectors of coefficients, one vector for each new dimension, that define the mapping function from the original coordinates to the coordinate value of a particular new dimension. The mapping function is the sum of the original coordinate values of a data point weighted by these coefficients. As a result, the eigenvalue ϵ_j indicates the amount of information in a new dimension j and the coefficients c_{ij} indicate the influence of the original dimension i on the new dimension j. Our PCA based ranking system (PCAr) makes use of these two facts by scoring the bands (the "original" dimensions in the above discussion) by (8).

$$PCAr(\lambda_i) = \sum_j |\epsilon_j c_{ij}| \quad (8)$$

As the procedure for computing the eigenvalues and coefficients is both complex and available in most data analysis texts [4], it is omitted.

H. Spectral Spacing

This method uses no information specific to the data set under consideration. Band are ranked so that for any set of top k bands, those k bands are as evenly spaced in terms of their central wavelengths as possible. For example, if 100 bands were to be ranked, their order would be $\{50, 1, 100, 25, 75, \ldots\}$. While this method may seem trivial, it actually takes into account a significant amount of domain specific-knowledge: bands that are near each other in the spectrum almost certainly contain similar information, bands that are far apart likely contain relatively unique information. From a data analysis point of view, incorporating such domain knowledge often can be more useful than any computed knowledge, no matter how sound the theory behind it may be.

IV. Supervised Classification Methods

A. Naïve Bayes

Bayes law (9) provides the posterior probability of an event C_i occuring given that event Λ has occured based on the the prior probabilities of C_i and Λ, as well as the posterior probability of event Λ given C_i. Here, this provides a means of calculating the probability of each possible class C_i given a spectral signature Λ and then selecting the class with the highest probability $P(C_i|\Lambda)$ as the prediction. $P(C_i)$ can easily be estimated from the set of training examples and $P(\Lambda)$, which is constant between classes, can be ignored as the classifier scheme is simply comparing the probabilities of different classes. To calculate the value of $P(\Lambda|C_i)$, conditional independence amongst attributes (here, spectral bands) is assumed (hence the name *Naïve* Bayes, which allows the use of (10).

$$P(C_i|\Lambda) = \frac{P(\Lambda|C_i)P(C_i)}{P(\Lambda)} \quad (9)$$

$$P(\Lambda|C_i) = \prod_k P(I(\lambda_k)|C_i) \tag{10}$$

In our implementation, the continuous variables $I(\lambda_k)$ are binned, and estimated probabilities based on training data are stored in a histogram for every $(C_i, I(\lambda_k))$ pair for use in (10). This introduces the need for control parameters for the binning method. The first parameter is a switch to select either binning by width or binning by depth. Binning by width takes a single interval size that all bins are given, with the lower bound of the first bin being the minimum value of the training set. In binning by depth, all bins are required to have an equal number of training examples, and the interval size is therefore variable between bins. The second parameter is therefore either the interval size or number of examples per bin, depending on which method is indicated by the first parameter. These parameters are optimized by the technique described in section V

B. Instance Based

Instance based classifiers, sometimes called k-nearest neighbors classifiers [11], [12], make a prediction for a test case based on the classes of the k training examples that have the smallest euclidean distance to that test case. The training stage of model building is therefore nothing more than storing the training examples. During prediction the distances to all n training examples must be calculated for each test case, and the k smallest (where k is a user defined control parameter) are selected. Often, the prediction is made by a simple majority-rules vote of these k nearest neighbors. Here, however, we bias the votes by the inverse of the distance to the test case, raised to the power w (another control parameter). This gives training examples with a smaller distance a higher weight in the voting. The weighted "vote" for each possible class C_i is therefore given by

$$V(C_i) = \sum_{e \in \{e:C(e)=C_i\}} \frac{1}{d_e^w} \tag{11}$$

where $C(e)$ is the class of training example e, and d_e is the euclidean distance $d_e = \sqrt{\sum_k \left(I(\lambda_k) - I(\lambda_k^e)\right)^2}$ from the training example to the test case in the spectral space. The number of neighbors k and the exponent weight w are optimized using the technique of section V.

C. C4.5 Decision Tree

A decision tree is a recursive search structure that can take on one of two forms: (1) a leaf, which has an associated class, or (2) a node that contains a test on a single attribute of the examples, and a branch and subtree for each possible outcome of that test [13].

C4.5 is widely considered the standard implementation of a classification decision tree. The learning process of a C4.5 decision tree involves finding the optimal test at each node to base the split on (or decide that the node should be a leaf). C4.5 exhaustively tries every reasonable test criterion at each node and selects the test based on some information gain criteria (see below). In the case of discrete attributes, this simply means creating a branch and subtree for every possible value of the attribute. For continuous attributes (the category spectral data falls into), C4.5 tries all $(m-1)$ possible values to perform a binary split for each attribute (less than evaluates to the left, greater than or equal to evaluates to the right), where m is the number of training examples that have evaulated to the node in question. Because all attributes are tested at each node, the algorithm can become quite expensive for large numbers of attributes.

The information gain indicates the decrease in variability of the classes in each of the subtrees. That is, it measures the uniformity of the class labels of the examples in the child nodes as compared to the parent. The information of a node, given in terms of the set T of training examples it contains, is given by:

$$H(T) = -\sum_j p(C_j|T) \ln p(C_j|T) \tag{12}$$

where the probability $p(C_j|T)$ is simply

$$p(C_j|T) = \frac{|\{e : e \in T, C(e) = C_j\}|}{|T|} \tag{13}$$

Finally, the information *gain* of a potential split S is given as the information of the parent minus the summation of the information content of its k children:

$$Gain(S) = H(T) - \sum_k \frac{|T_k|}{|T|} H(T_k) \tag{14}$$

Where T_k is a set of examples that is the subset of T that evaluate to the same child node. The potential split with the highest gain is selected and the algorithm is repeated on the children. A node is declared to be a leaf if either a minimum information gain threshold τ_i is not satisfied by the best potential split, or similarly if the number of training examples in the node is less than the minimum examples per leaf τ_e. Both τ_i and τ_e are user defined parameters that are optimized by the method from section V.

V. Experimental Verification

We implemented our ROWAS procedure as follows. First, the bands are ordered by the unsupervised method. A set of bands is initialized with the top two ranked bands, and at each iteration two additional bands are added to the set. For each of these sets, one hundred random sets of control parameters for each supervised method are tested, and the best set of parameters is determined using eight fold cross-validation. The process of n-fold cross validation is used to test the performance of a model given a single data set. The data is split into n subsets of examples, and n models are constructed using each subset as a hold-out set. The models' accuracy is scored on the hold-out set for each respective model, and the average accuracy is accepted as the accuracy that the model can achieve in the domain the dataset is derived from. The final

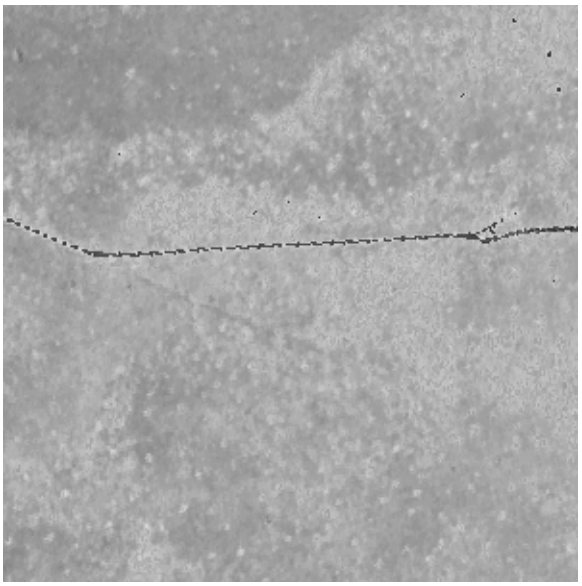

Fig. 1. Grayscale composite of the bands from 1330nm to 2040nm of the experimental data. Taken October 20, 1999

error for a set of bands and supervised method combination is determined by a final twelve fold cross-validation using the parameters determined in the random optimization step. This entire process is then repeated for each unsupervised ranking method.

To test our procedure, we obtained a data set that consisted of spectral measurements from an AVIRIS [14] sensor and manually collected labels of the grass type of scanned regions [15]. The AVIRIS sensor is a whiskbroom type sensor with a spectral response of 400 to 2500 nm, with 224 contiguous channels, approximately 10 nm wide. The spatial response was 0.87 mrad, which translates to approximately 3.2×3.2 m pixels for readings taken from 1700 ft (the altitude our test image was taken from). The set of ground labels was $\{$Unclassified, Black Grama, Blue Grama, Road, Black Grama/Green Veg Mixed, Blue Grama/Green Veg Mixed$\}$. Fig. 1 shows a single band of the spectral image, which was taken October 20, 1999.

A. Results

The top score for each supervised, unsupervised method pair is given in Table I. The score is the sample mean absolute error obtained from the final 12-fold cross-validation performed for every set of top ranked features. Also given is the number of bands used to achieve the best score (denoted as 'count'), which indicates how effective the unsupervised method was at selecting the best bands first.

The graphs of figures 2 - 4 show the complete results for the three unsupervised methods with the best scores for each supervised method. Also included are a random ranking and an average random plot. In addition to the rankings generated by the supervised methods, six random rankings were tested using the same framework. The random ranking that performed best, as well as the average over the random trials, correspond to these two plots, respectively. These random rankings provide a baseline for comparison. The other unsupervised and random rankings are omitted for the sake of clarity in the graphs.

TABLE I

THE NUMBER (COUNT) OF TOP RANKED BANDS USED TO ACHIEVE THE BEST SAMPLE MEAN ABSOLUTE ERROR, AND THE ERROR ITSELF.

	Naïve Bayes		Instance Based		Decision Tree	
	Error	Count	Error	Count	Error	Count
Entropy	.068	92	.024	38	.081	18
1^{st} Deriv.	.105	64	.040	42	.049	22
2^{nd} Deriv.	.105	24	.040	96	.053	48
Contrast	.064	98	.032	42	.085	16
Ratio	.113	14	.028	98	.049	18
Correlation	.081	90	.045	52	.061	86
PCAr	.065	68	.024	46	.117	42
Spectral Spacing	.061	20	.020	62	.113	60
Best Random	.048	10	.016	24	.081	8
Average Random	.063	36	.026	76	.116	92

Naïve Bayes (Fig. 2) does the least well as a supervised method. This is not totally unexpected, as it makes the strong assumption of conditional independence among the input features. The spectral information, however, is highly correlated, especially among bands near each other in the spectrum. Also noteworthy in Fig. 2 is that the performance seems to be asymptotic as the number of bands grows. Because the different bands contain similar information, and because of the nature of the algorithm that treats all bands equally, it's not unlikely that the additional bands are simply smoothing out the noise inherent in the data set and also the noise generated when the data is binned.

The spectral spacing and contrast methods do the best out of the intelligent methods. Because of the issue of correlated features near each other in the spectrum, using the most spread out bands for any given set of bands should cause the fewest problems (although it doesn't address the issue of whether those bands are actually *relevant*). This is exactly what the spacing method does. The contrast method does the second best, but the optimum is not reached until 94 bands are used. For our purposes, this makes it little better than any other method, as all show asymptotic behavior and an optimum using so many bands proves little about the suitability of the ranking method for this domain. This is compounded by the fact that the *average* random optimum was superior to all of the supervised methods except spectral spacing. The supervised methods therefore are not considering the information relevant to achieving high accuracy with a naïve bayes classifier. Furthermore, the best random ranking beat even the spectral spacing method. This ranking likely ordered bands in such a way that they were not only reasonably uncorrelated, but also had high information content in the top ranks.

Next was the instance based classifier, with the best results shown in Fig. 3. Instance based classifiers can be finely tuned to a data set due to its parameters that can vary the behaviour

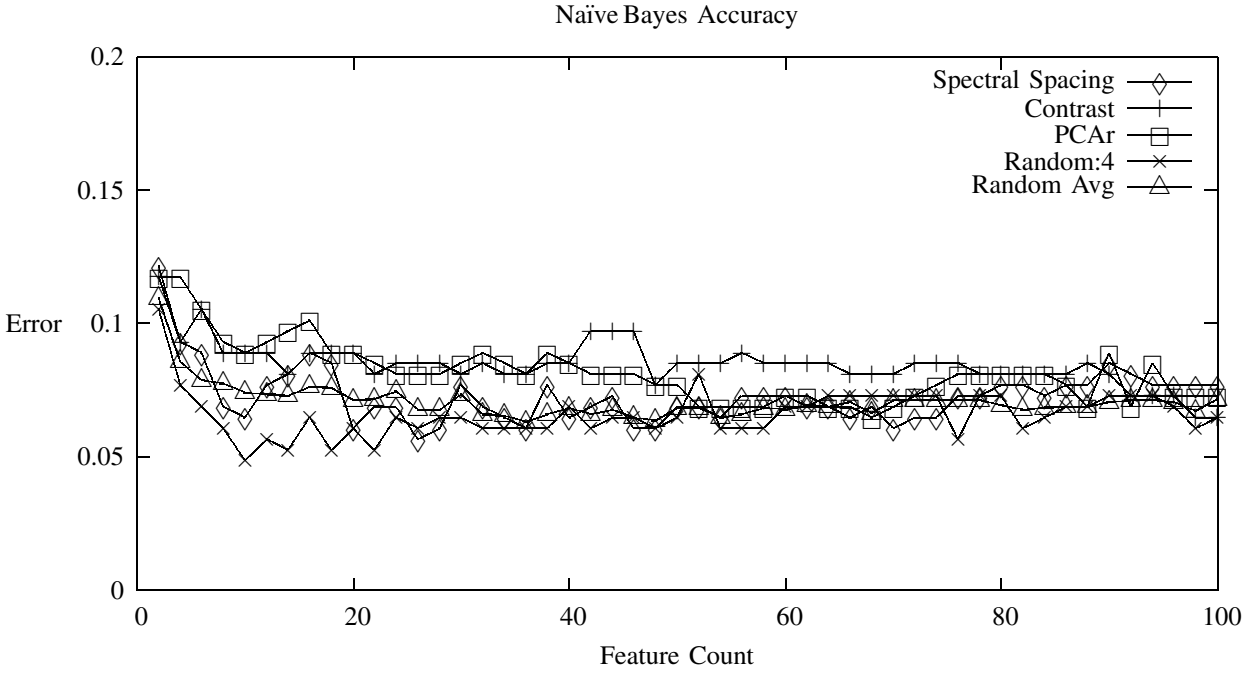

Fig. 2. Naïve Bayes accuracy for top unsupervised ranking methods.

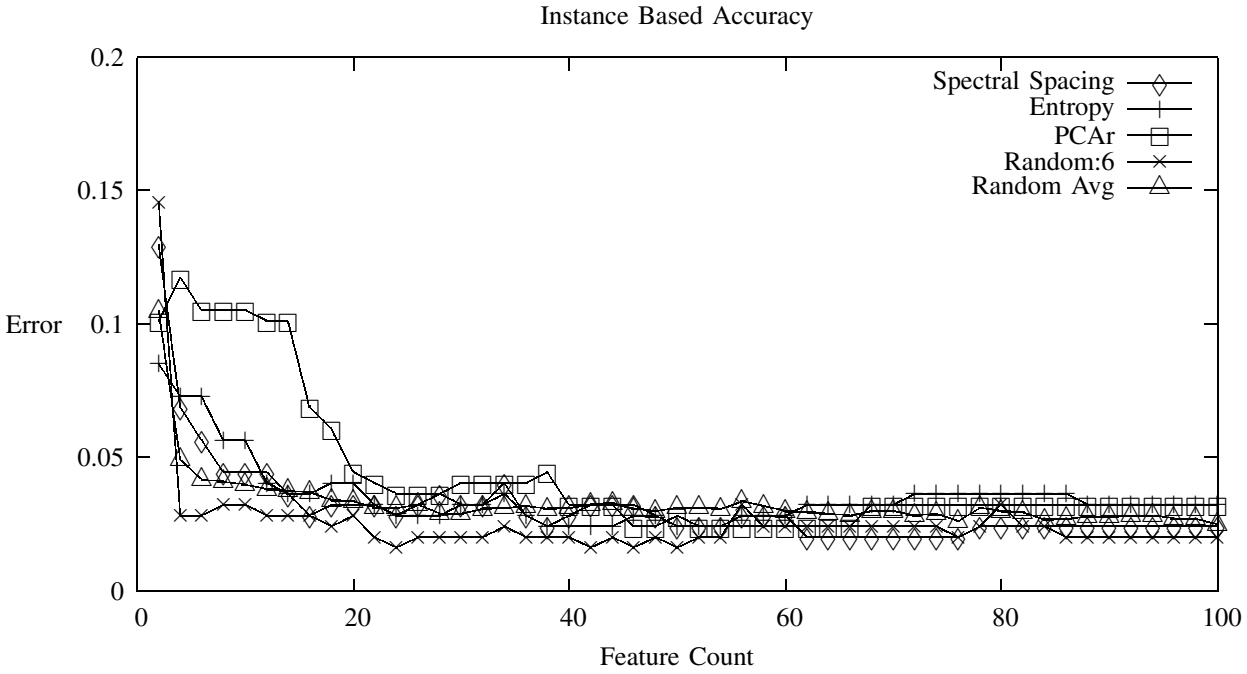

Fig. 3. Instance Based accuracy for top unsupervised ranking methods.

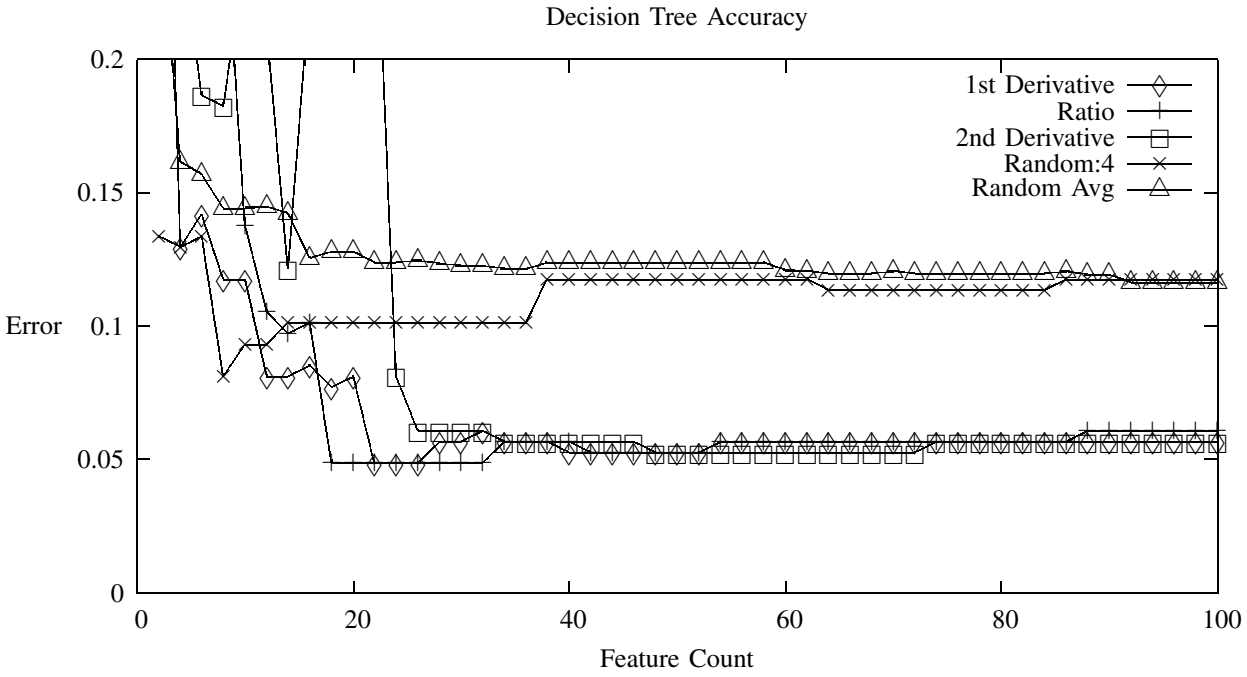

Fig. 4. Decision Tree accuracy for top unsupervised ranking methods.

of the classifier greatly. While slower than naïve bayes, it typically performances at least comparably, and often better. Its accuracy depends not only on the parameters, but also on the relevance of the feature set. Irrelevant features are given as much weight as relevant ones, and simply add noise to the predictions. Redundant features can give too much weight to some information at the expense of that found in other features. This was verified by the fact that the entropy and PCAr methods performed well, as they both produce rankings based on the information content of the bands. They do not take into account the redundancy of bands, and are inferior to the spectral spacing method. The optimal point for the spacing method came at 62 bands, while the entropy and PCAr methods performed best at 38 and 40 bands, respectively. Furthermore, the difference between the scores was merely 0.004, which means that only a single example more was classified correctly by the optimal point of the spacing method. The best random ranking performed another 0.004 better. This, again, shows that our unsupervised methods do not produce optimal rankings, but that the efficiency of our overall procedure allows enough methods to be evaluated to detect such deficiencies. A further discussion of the significance of these differences follows in (V-B).

Finally, the C4.5 decision tree results are given in Fig.4. Decision trees do their own greedy search over features that have the most impact on the information content of the *predicted* variable. While they normally perform well, they can suffer if noisy or irrelevant features lead them astray early in the tree building process. Furthermore, having very similar information in two features has been known to degrade performance, as the decision for which of two similar features to use is determined primarily by noise (the noisier feature may well be picked). Somewhat surprisingly, spectral ratio and 1^{st} and 2^{nd} derivative ranking methods did by far the best. These methods performed rather poorly with the other supervised methods. They even had only half the error rate of the spectral spacing method, and nearly half the error rate of the best random ranking (error rates of 0.049 for ratio and 1^{st} deriv., 0.117 for spectral spacing, and 0.081 for the best random). These three ranking methods work by comparing every band to a band immediately adjacent to it in the spectrum, promoting those band pairs that exhibit less correlation. It's likely this allowed these two ranking methods to overcome the problem of poor performance due to similar features.

B. Significance Test

Due to the small differences in accuracy among the top ranking methods for the respective classifiers, we employed a statistical significance test to determine the confidence level of the superiority of the best methods. The McNemar test for categorical/nominal data of dependent samples [16] was used. Dependent samples, which are normally used in the social sciences, involve using the same test subjects with different treatments, and measuring a boolean response after each treatment. The two treatments are then compared to determine if one was more likely to cause the response than the other. In our experiment, this equates to the same test example being classified with two different classifiers. The boolean response is then whether the classification was correct.

The McNemar test compares two sets of predictions, A

TABLE II

THE RESULTS OF THE MCNEMAR SIGNIFICANCE TEST BETWEEN THE BEST SETS OF PREDICTIONS FOR EVERY SUPERVISED METHOD AND OTHER (COMPARED) TOP METHODS. THE CONFIDENCE LEVEL IS THE LIKELIHOOD OF THE NULL HYPOTHESIS STATING THAT THE TWO UNSUPERVISED METHODS GENERATE THE SAME POPULATION OF PREDICTIONS.

Supervised Method	Best Unsupervised Method		Compared Unsupervised Method		Confidence
	Name	Count	Name	Count	%
Naïve Bayes	Random:4	10	Spectral Spacing	20	29.05
Naïve Bayes	Random:4	10	PCAr	68	22.72
Instance Based	Random:6	24	Spectral Spacing	62	50.0
Instance Based	Random:6	24	PCAr	46	25.00
Instance Based	Random:6	24	Entropy	38	25.00
Decision Tree	Ratio	18	1^{st} Derivative	22	100.00
Decision Tree	Ratio	18	2^{nd} Derivative	48	50.00
Decision Tree	Ratio	18	Random:4	8	59.82

and B, as follows. The number of test examples that classify correctly for one prediction set, but not the other, are tallied for both sets of predictions. If we assume that the two sets of boolean right/wrong values come from the same distribution (because the sets of predictions are from the same distribution), then it follows that there is a $\pi_a = 0.5$ and $\pi_b = 0.5$ probability that prediction set A or B will be the correct one for any given test example. That is, if only the examples that evaluate differently are considered, then the results would be equally distributed if they came from the same distribution. Using the binomial distribution, the likelihood of obtaining the observed tallies is computed by (15).

$$P(\geq x) = \sum_{r=x}^{m} \binom{m}{r} (\pi_b)^r (\pi_a)^{(m-r)} \tag{15}$$

Where x is the tally for the more accurate prediction set, m is the sum of the two tallies, and $\pi_a = \pi_b = 0.5$ are the probabilities that one prediction set will be correct when the other is not. The value obtained from (15) is the probability of obtaining the observed predictions if the two prediction sets were drawn from the same population. The assumption that $\pi_a = \pi_b$ is therefore the null hypothesis, and (15) is the confidence that it is true. A lower value therefore means it is more likely that the ranking method with a higher accuracy was truly better than the one it is being compared to.

There are two important considerations when using the McNemar test. First, it only takes into account those test examples where the predictions were different. The test does not rely on the total number of samples in any way. The second consideration follows from the first: the test ignores both examples where both predictions are correct *and those where both are incorrect*.

Confidence levels that the top ranking method was statistically the same as the next best methods are given for every supervised method in Table II. Again, the lower the confidence level, the more likely the best method is statistically superior.

When employing a significance test, it is common to require either a 95% or 99% confidence level to accept a hypothesis. Therefore, to accept the hypothesis that the best ranking for a supervised method is truly better than the second best rankings, the value in Table II must fall below at most 5%. Not surprising because of the small differences in accuracies, this never occurs. In all cases except ratio and 1^{st} derivative with decision tree, it can not be said with much certainty that the two prediction sets *are* likely to be from the same population, either. If we assume that one of the rankings tested (even if it is one of the random rankings) is at or near the theoretical optimal accuracy for this data set, we can conclude that the top unsupervised methods are performing at a level insignificantly below that optimum.

VI. CONCLUSION AND FUTURE WORK

We have presented a method for feature set selection and classifier method selection that makes few assumptions and performs well while using a significant, but tractable, amount of computing power. The evaluation of top ranked bands exhibited definite trends that allowed for the discovery of the optimal number of bands.

In our empirical study, we found that for the AVIRIS image with gramma grass labels we used the instance based classifier performed the best. While one of six random rankings performed the best overall, the spectral spacing, entropy, and PCAr methods did very well, as well. The difference in error rates between the best random and best unsupervised methods translated into at most two additional examples of the 247 examples being classified correctly (six incorrect classifications for PCAr and entropy, five for spectral spacing, and four for the best random). Upon further analysis, we concluded that these differences were not statistically significant. Furthermore, if we make the (admittedly unsupported) assumption that the best feature subsets produced by the ranking methods are indeed optimal, we can conclude that at least one of the unsupervised methods was able to reliably produce results only marginally inferior (but not significantly inferior) to the optimal. This is an important point to emphasize, as random rankings had the highest overall accuracy for the instance based and naïve bayes classifiers. There were other random rankings, however, that had significantly *worse* performance than the best intelligent methods. We can then say that we are able to get excellent

performance with relatively little computing power, as opposed to requiring the testing of many random rankings in search of a stand-out winner.

The ROWAS method exhibited easily explainable behaviour and produced high classification accuracy. While some of our unsupervised methods did not perform as well as hoped, the efficiency of the evaluation mechanism allowed them to be quickly and definitively identified.

As our method relies on self-contained components (unsupervised methods, supervised methods, supervised method parameter optimization), there are numerous possibiltities for future work. First, ranking techniques that balance the tradeoff between information content and redundancy, instead of emphasizing one or the other, should make the most immediate improvement. The naïve bayes classifier can also be safely replaced by another supervised method. Because of the stability of the bayesian classifier we used, we can conjecture that bayesian methods do model the given problem correctly in general, the one we chose was simply not the optimal one. Another bayesian classifier that does not make such strong assumptions about independence (which were known to be invalid here), such as a bayesian belief network, may therefore be a worthwile replacement.

REFERENCES

[1] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inform. Theory*, vol. IT-14, no. 1, January 1968.

[2] J. C. Price, "Band selection procedure for multispectral scanners," *Applied Optics*, vol. 33, no. 15, pp. 3281–3288, 1994.

[3] J. A. Benediktsson, J. R. Sveinsson, and K. Arnason, "Classification and feature extraction of AVIRIS data," *IEEE Trans. Geosci. Remote Sensing*, vol. 33, pp. 1194–1205, 1995.

[4] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Recognition, Second Edition*. New York: Wiley-Interscience, 2000.

[5] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997. [Online]. Available: citeseer.nj.nec.com/article/kohavi97wrappers.html

[6] J. C. Russ, *The Image Processing Handbook*, 3rd ed. CRC Press, 1999.

[7] D. J. Wiersma and D. A. Landgrebe, "Analytical design of multispectral sensors," *IEEE Trans. Geosci. Remote Sensing*, vol. 18, no. 2, pp. 180–189, 1980.

[8] J. B. Campbell, *Introduction to Remote Sensing*, 2nd ed. New York: The Guilford Press, 1996.

[9] T. Warner, K. Steinmaus, and H. Foote, "An evaluation of spatial autocorrelation-based feature selection," *International Journal of Remote Sensing*, vol. 20, no. 8, pp. 1601–1616, 1999.

[10] X. Jia and J. A. Richards, "Efficient maximum likelihood classification for imaging spectrometer data sets," *IEEE Trans. Geosci. Remote Sensing*, vol. 32, no. 2, pp. 274–281, March 1994.

[11] L. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers, 2000.

[12] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Francisco, California: Morgan Kaufmann Publishers, 2001.

[13] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco: Morgan Kaufmann Publishers, 1993.

[14] G. Vane, "Airborne visible/infrared imaging spectrometer (AVIRIS): Description of the sensor, ground data processing facility, laboratory calibration, and first results," Presented at Imaging Spectroscopy II Conference, San Diego, California, August 1987, Jet Propulsion Lab, NASA, Pasedena, California, Tech. Rep. JPL Publication 87-38, November 1987. [Online]. Available: http://aviris.jpl.nasa.gov/

[15] C. Wessman. (1999, October) Hyperspectral imagery with gramma labels. CU-CIRES-CSES, NASA Airborne Science Program, AVIRIS Project, Sevilleta LTER Database. [Online]. Available: http://sevilleta.unm.edu/data/archive/synopses/RS_datasets_update.html

[16] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures, Second Edition*. Boca Raton, FL: Champman & Hall/CRC, 2000.