RANK ORDERING WITH ACCURACY SELECTION (ROWAS) FOR HYPERSPECTRAL BAND SELECTION

BY

PETER GROVES

B.S., University of Illnois at Urbana-Champaign, 2001

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Masters of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2003

Urbana, Illinois

RANK ORDERING WITH ACCURACY SELECTION (ROWAS) FOR HYPERSPECTRAL BAND SELECTION

Peter Groves, M.S.
Department of Computer Science
University of Illinois at Urbana-Champaign, 2003
Peter Bajcsy, Advisor

Hyperspectral band selection is a key factor in creating practical, accurate predictive models for remote sensing applications. A proper subset of bands can contain the same information, with less noise, than the complete set of bands. This can lead to both an increase in accuracy and a decrease in computational complexity. The problem then becomes: how does one determine which bands to use? We first discuss the implications of sampling theory, the No Free Lunch Theorem, sensor noise, and information redundancy in feature subset selection. We then present a generic methodology that directly follows from these implications to select the optimal subset of bands and prediction model together. This method is called Rank Ordering with Accuracy Selection (ROWAS), and works as follows. The bands are ranked by several computationally efficient measures of information content and redundancy. Then, increasing numbers of top ranked bands are evaluated with different prediction methods using the cross-validated accuracy as a metric. The ensuing analysis provides an optimal set of bands, along with the best prediction model. This methodology satisfies all of the design constraints, and provides a good tradeoff between exploration of the feature space and computation time.

We apply this generic methodology to the domain of hyperspectral band selection by developing ranking methods that assume the data is a sampled continuous spectrum. Experimental results for both a numeric prediction and classification task are presented. These experiments are the the prediction of electrical soil conductivity in a pre-growing season farm field and the classification of grass types based on hyperspectral, airborne imagery. For both problems, ROWAS achieves a high level

of accuracy appears to be near the optimal accuracy possible for the problem. In the case of the grass-type classification, this is confirmed using a McNemar test for statistical significance.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Remote sensing provides a means to compile data with varying spectral and spatial resolution over large geographic areas by mounting sensors from airplanes or satellites. The resulting spectral data has a wide range of applications in environmental monitoring [1, 2, 3, 4], sensor design [5, 6], geological exploration [7, 8], agriculture [9], forestry [10], security [11], cartography, and the military [12, 13]. In nearly all applications, the underlying problem is the desire to use spectral information to predict certain properties of the objects being imaged. This can involve predicting a categorical/discrete variable, often termed classification, or predicting a continuous variable, known as regression.

To provide greater spectral resolution, hyperspectral imagery is often used to model a remotely sensed scene. In hyperspectral imagery, the electromagnetic spectrum is sampled at tens, hundreds or even thousands of wavelength ranges in the visible and near infrared (NIR) specta, hereafter called *bands*. The result is a very detailed view of the spectral signature of the scene represented by a particular pixel. The additional information comes at a cost, however. With more features to use for prediction comes additional noise, redundancy, and model complexity that can degrade accuracy [14]. Practical considerations such as computation time, storage, and communication bandwidth must also be acknowledged for the sake of end user

applications.

Problem Statement

Find the subset of hyperpectral bands that efficiently maximizes the accuracy of a predictive model using an algorithm that is efficient in terms of both computation time and user effort.

This thesis focuses on the role of feature selection (band selection) in hyperspectral image classification. A method for balancing the tradeoffs between exploration of feature subset search space and computation time is proposed. Theoretical and practical considerations are discussed, and experimental results for both classification and regression are presented.

1.1 Prediction Models

In this thesis we define a *prediction model* to be an empirically derived function that provides a one-to-one or many-to-one mapping from a set of input features to a set of predicted output values. Both input and outputs can be categorical, nominal, or continuous variables. Here we are mainly concerned with the case of continuous input features, as that is the form of the hyperspectral data. To provide some degree of robustness, however, we will explore the prediction of both continuous and categorical data. The general framework we will develop can be applied to data of any form.

In the field of remote sensing, models can generally be put into one of two categories: pixel based classifiers and spectral-spatial classifiers [15]. Pixel based classifiers assume that the output variable is a function solely of the spectral information contained in a pixel. Spectral-spatial classifiers use the spectral information as well as the spectral information of surrounding pixels, giving the model a degree of spatial, or contextual, response. Here we focus on pixel based classifiers as they have the form of

most statistical or machine learning modeling techniques. This form is characterized by having a set of features that are thought to have equal relevance *a priori* to the model building process and have no direct interaction between them. That is, there are no features that are calculated as functions of other features, although this does not mean they are not correlated or statistically dependent.

1.2 Previous Work

As the amount of data being collected increases at an exceptional rate feature subset selection has been a key subject in many areas that use data mining and modeling. There is therefore a large body of work of both general methods and those related specifically to hyperspectral data. We will briefly present some of these methods by categorizing them first as either top-down or bottom-up, and then as either filter or wrapper methods.

The basic heuristic of many feature selection techniques can be called top-down or bottom-up selection schemes [14]. Top-down algorithms start with the entire set of features and iteratively remove the worst feature(s). Bottom-up methods work in reverse, starting with no features and adding the best. What determines the best or worst feature is determined by the particular implementation. Often a simple statistical significance test such as correlation is used. That is, a good feature will have little correlation with the other input features and high correlation with the predicted variable.

A somewhat more sophisticated measure is used by the *wrapper* method, formalised by Kohavi and John [16]. The Wrapper Method scores a subset of features based on the cross-validation accuracy of a prediction model. This allows it to be used with greedy top-down and bottom-up methods [17], as well as more robust optimizers such as genetic algorithms. Yang and Honavar, for example, proposed a genetic algo-

rithm, neural network wrapper combination that provided high classification accuracy [18].

Selection techniques that do not consider the prediction model (as wrapper methods do) are often termed *filter* methods. This can include methods such as the aforementioned correlation tests. A flaw common to many filter techniques is the tendency to throw out *both* features if two are found to contain the same information. Kohavi and John explored this topic in [17] and concluded that wrapper methods are therefore generally superior to filter methods.

Recently, Tsamardinos and Aliferis refuted that claim based on the No Free Lunch Theorem (NFL) [19]. The NFL applies to any optimization algorithm. At its core, it states that no optimization algorithm is better than another when the performance is averaged over all possible problems [20]. The NFL theorem is an integral part of the methodology developed in this thesis, and will be discussed further in Section 2.2. Tsamardinos and Aliferis concluded that because of the NFL theorem, filters can be as accurate as wrapper methods, provided they take into account the prediction model and error metric being used.

1.2.1 Hyperspectral Band Selection

Because of the many applications of remote sensing, considerable work has been done in hyperspectral feature selection. Such work may deviate from general feature selection methods because of the specific structure of hyperspectral data. Namely, bands (features) that are near each other in the spectrum are likely to share certain properties and be highly correlated.

Closely related to feature selection is feature *extraction*. In feature extraction new features are generated that contain the same amount of information as the original feature set, but in a lower dimension space. Examples of this are Principle Component Analysis (PCA), Independent Component Analysis (ICA), and wavelet transforms

[21]. In this work we have avoided feature extraction for two reasons. One is that it is difficult to interpret the results of an analysis based on extracted features. Another is that they make a prediction system less generic, as they can rely on certain assumptions about the structure of the data that we are trying to avoid.

While a few band selection techniques have been proposed, such as the ICA based selection procedure described in [22], they are far less common than extraction methods.

1.3 Methodology Overview

To overcome the obstacles mentioned above (which will be elaborated on in Section 2.3), we propose a new method: Rank Ordering with Accuracy Selection, or ROWAS. The basic structure of the methodology is simple. First use computationally efficient unsupervised methods to rank order bands based on their information content and distinctiveness. Then, use a wrapper approach with a supervised prediction algorithm to evaluate sets of top ranked bands. The experimentally determined optimal accuracy will reveal the prediction model, band suitability measure, and number of bands that are best suited for a given problem. A variety of unsupervised and supervised methods are used to remove the need for *a priori* knowledge of the underlying structure of the data. We will in this study, however, use several unsupervised methods that do take into account the spectral relationship of the input features.

1.4 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 3 formalizes the considerations a modern feature extraction system must resolve, and discusses how ROWAS meets those challenges. Chapter 4 gives a detailed listing of the unsupervised and supervised methods we propose for use in a ROWAS system meant for use in the

hyperspectral domain. Two experimental data sets, one with a categorical predicted variable and one with a continuous predicted variable are analyzed using ROWAS in Chapter 5. Finally, a summary of the results and our conclusions are presented in Chapter 6.

Chapter 2

Rank Ordering With Accuracy Selection: A Feature Selection Methodology

2.1 Feature Selection

The problem of feature selection is not unique to the field of hyperspectral remote sensing. The related fields of statistical modeling, pattern recognition, and machine learning share the fundamental problems caused by improper feature selection. The four main concerns that have driven the field of feature selection are computation time, redundancy, irrelevance, and the so-called Hughes Phenomenon [23]. A brief discussion of each is given below.

In building a predictive model through machine learning techniques, the number of possible models increases exponentially with the number of features. That is not to say training methods will consider all possible models, but this indicates that the search space for the appropriate model is much larger. Take, for instance, a simple decision tree. The greedy tree building mechanism normally employed causes the building time for a binary tree to be roughly $O(mn^2(\log n))$ where m is the number

of features and n is the number of training examples [20]. This is a heuristic to find a good solution to what is actually an NP-hard problem [24] (NP-hard problems can loosely be defined as those that take exponential time to solve). Therefore, while the actual computation time will increase only linearly with additional features, the search space will increase exponentially, making it more difficult for the greedy hill-climbing style algorithm to find the global optimum. This reasoning applies to other modeling algorithms that partition the feature space directly, such as support vector machines (SVM), as well.

2.1.1 Noise and Irrelevance

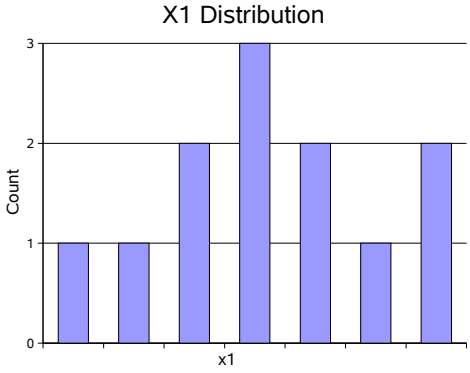
While the Hughes Phenomenon of Section 2.1.2 and the sampling considerations of 2.1.3 provide arguments for the theoretical optimal size for a subset of features, two issues that could possibly dominate the problem have not been mentioned. They are noise and irrelevance. If a feature has a very poor signal-to-noise ratio, it is apparent that the ability to use that feature in a modeling situation will be degraded. Likewise, if a feature is completely unrelated to the variable being predicted, the information contained in that feature will appear the same as noise to the prediction model.

2.1.2 Hughes Phenomenon

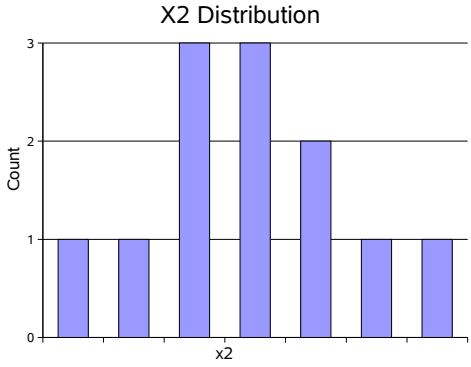
Related to the problem of an exponential increase in the size of the model search space (given above) is what has become known as the Hughes Phenomenon, based on the findings of Gordon Hughes in 1968 [23]. Fundamentally, the Hughes Phenomenon is that the density of a fixed amount of data points decreases as the number of dimensions increases, inhibiting the ability to make reliable estimates of the probability distribution. The result is that there is an optimal number of features to use for a given number of data points (assuming they are all valid and dependent on the predicted class).

Table 2.1: Example data to demonstrate the Hughes Phenomenon. One and two dimensional histograms are given in Figure 2.1.

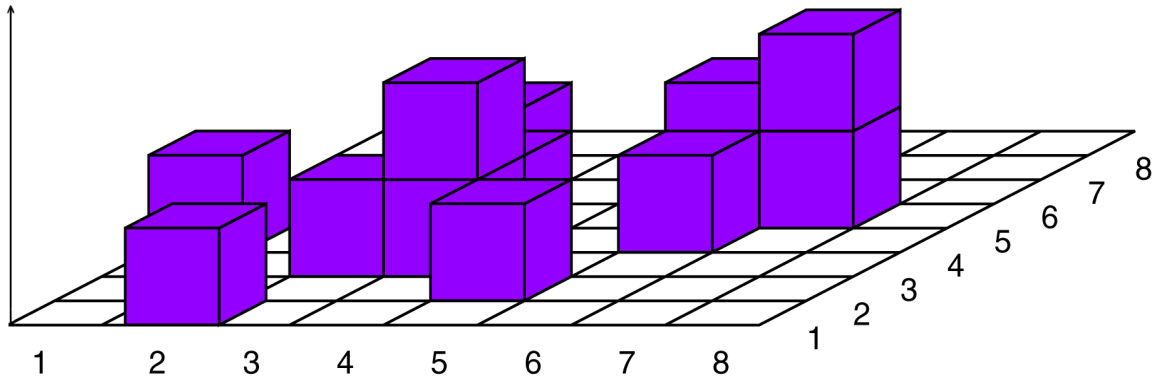
X1	1	2	3	3	4	4	4	5	5	6	7	7
X2	4	1	6	3	3	4	3	2	7	4	5	5



(a) A histogram showing the sample of X1 from Table 2.1.



(b) A histogram showing the sample of X2 from Table 2.1



(c) A histogram of X1 and X2 together. Note that little can be determined about the distribution of the underlying distribution given the small sample size.

Figure 2.1: A demonstration of the Hughes Phenomenon. Estimates of the distributions of the variables given in Table 2.1 can be made based on the univariate histograms, but not on the bivariate histogram.

A simple example is given in Figure 2.1. One dimensional histograms of the data given in Table 2.1 are shown in Figures 2.1(a) - 2.1(b). With the data given, the trends of the one-dimension probability distribution functions (PDF's) have begun to emerge and the parameters of a Gaussian distribution could be estimated with some confidence. The same is not true of the joint probability distribution shown in Figure 2.1(c). The same number of data points in a larger dimension space appear more sparse. Even if a multi-variate Gaussian model was assumed, the parameters estimated from the given data would not be very trustworthy. As the number of dimensions increases with a fixed number of data points, it is easy to imagine how the problem would simply get worse.

2.1.3 Implications of the Central Limit Theorem

While the Hughes Phenomenon shows how even correct data can be misleading if there is not enough data to make density estimates in a high dimension space, there is also an argument for collecting data with as many dimensions as possible, but only using a subset of those dimensions.

The Central Limit Theorem is one of the foundations of modern statistics. It states that for samples of a given size, taken from identical distributions, the sample means will themselves converge to a normal distribution, even if the sampled population has some non-normal form. What is more, the variance of the means over the different samples is a function on the size of the sample, as is how closely the distribution can be approximated with a normal distribution. That is, as the sample size increases, the distribution of sample means becomes more normal and has a smaller variance [25].

This has important implications for the problem of feature subset selection. Assume we have a fixed data set size of n samples and m features. Let $X_1, X_2, ..., X_m$ be the sample sets of size n. If for all $\{i : 1 < i \leq m\}$ the (true) population distri-

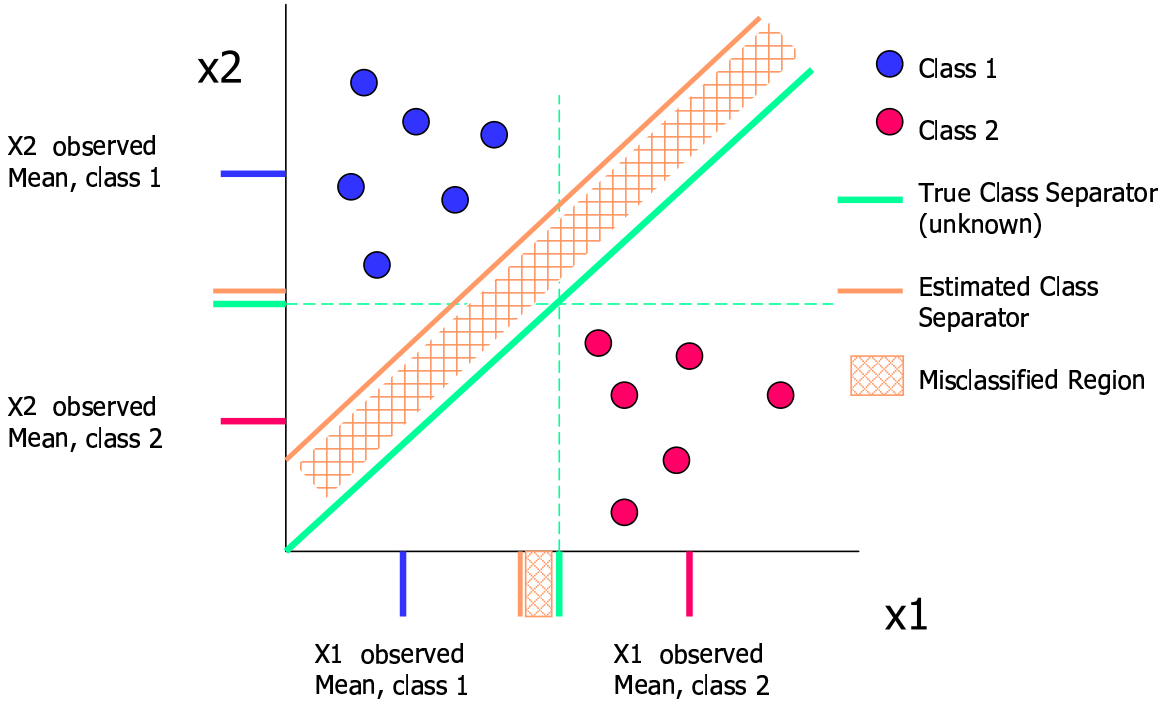


Figure 2.2: An example of sampling errors leading to poor accuracy. Note that the misclassified region is larger when both dimensions are used, as opposed to only $X2$. Also, the red points all fall within the true class boundaries, the sample mean is simply different than the population mean, mainly due to insufficient sample size.

butions have mean μ_i and variance σ_i^2, then $E(\bar{X}_i) = \mu_i$ and $V(\bar{X}_i) = \sigma_i^2/n$ [25]. We therefore have a situation where any sampled set X_i has a mean that is most likely to be its true population mean, but with a variance proportional to σ_i^2 and inversely proportional to (the constant value) n. We can therefore conservatively conclude that out of the m features, "some" of the samples will have estimated parameters (namely, mean) that more closely represent their true population parameters than others. If we set a threshold η to accept or reject a feature's sample as adequate by the rule $\eta < |\mu_i - \bar{X}_i|$, it is obvious that the expected number of features that satisfy the rule will increase as the number of total features m increases.

A demonstration of how this could affect a classifier is given in Figure 2.1.3. A simple binary classifier is shown that can classify by one of three metrics: minimum distance to the class mean in the $X1$ dimension, minimum distance to the class mean in the $X2$ dimension, and minimum distance to the class mean in both dimensions.

Both the true boundaries and the boundaries that would be calculated from the observed data are shown. Note that because the observed mean for $X1$ does not match the true class mean for the red class, there is a significant region of misclassified points when using either $X1$ alone or $X1$ and $X2$. Furthermore, all of the red points fall in the true class boundaries, and are therefore not excessively noisy. The probability that the misclassification regions would be reduced would increase if more red data points were included.

It should be noted that the effects of sampling errors will becoming insignificant as the size of the data set n becomes substantial. However, we are mainly concerned with applications where data are expensive, and we cannot require many labeled points in our algorithms, as we are unlikely to get them. For instance, why would we try to predict features on the ground using a satellite if it was easy to determine the features through direct measurement? The very problem is that we do not have many labeled points and would like a new, cheaper way to label more.

Unfortunately, as the population means μ_i are unknown, we can not use the above thresholding technique directly for feature selection. The analysis does tell us something about what a feature selection algorithm must consider, however. For any two data sets, even if they are sampled from the same population, the features that more accurately represent the true population will likely change from sample to sample.

If we combine this conclusion with the Hughes Phenomenon, we see that for a given sample size, there is a fixed limit for an ideal number of features k for achieving optimal model accuracy, and a high likelihood that the features will have an ordering from most representative of their populations to (admittedly marginally) unrepresentative of those populations. We believe that this provides further evidence that feature selection from a large set of features is superior to feature extraction or dimensionality reduction techniques because poor features will influence the final data

set in the latter but not in the former.

2.2 No Free Lunch Theorem

As mentioned in Section 1.2, the No Free Lunch Theorem (NFL) states that the accuracy of any two optimization algorithms are equal when averaged over the set of all possible problems [20]. This has important consequences not just for feature selction, but also predictive modeling in general. In order to have a robust feature selection and modeling *system*, that system must be able to deal with the reality that the best model may change from one domain to another, just as the best feature selection scheme may change. Furthermore, the best feature selection scheme may depend not only the domain, but on which model is best.

Traditionally, the NFL theorem has led to the doctrine that a domain expert must analyze the problem as much as possible in order to intelligently prune the search space of potential optimizers. We feel, however, that as the computation power of modern desktop machines increases, and the advancement of distributed and parallel computing techniques continues, that the cost of additional computations will pale in comparison to the cost of a domain expert's time.

2.3 ROWAS

With the problems and issues surrounding feature selection defined, we are ready to present our proposed methodology. First, let us summarize the considerations we wish to resolve.

Hughes Phenomenon As the number of dimensions increases, the density of a fixed number of data points decreases in the space. This makes it more difficult to reliably estimate density functions in higher-dimensional spaces. For a given

problem with a set number of data points, there is an optimal number of features to use in a predictive model. (Section 2.1.2)

Sampling Reliability The distributions of samples will vary from their true populations by known probabilistic rules, defined by the Central Limit Theorem. As the number of features to choose from increases, the likelihood of being able to find a good subset of a certain size increases. (Section 2.1.3)

Irrelevance If a feature is unrelated to the variable being predicted, that feature is no different than noise, and can only degrade accuracy. (Section 2.1.1)

Noise Regardless of the cause of the noise, a model trained on noisy data will not be able to correctly make generalizations about the relationship between input features and predicted features.(Section 2.1.1)

No Free Lunch There is no optimization technique that is superior in all domains. Without prior domain knowledge, there is no reason to prefer one over another. This has two implications here:

- The best prediction model can not be determined in advance
- The best feature selection technique will depend on the prediction model used.

(Section 2.2)

Computer Vs. Expert's Time A domain expert's time is valuable, and will remain valuable. A computer's time is continually becoming cheaper. There is little reason to use a domain expert's time for a job a computer can do, so long as the computer does not take considerably longer. (Section 2.2)

2.3.1 Proposed Methodology

To overcome the above restrictions, we propose to rank order the hyperspectral bands by various metrics of information content, and then evaluate the prediction error of the i top bands at a time using different prediction models, 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). The error is considered to be the lowest error obtained when the parameters of the prediction model are optimized. For instance, in the k-nearest neighbors algorithm, k is optimized. We again leave the actually optimization process as an abstraction (in our experiments we will simply use a random optimizer; one that tries 100 random parameter sets and selects the best). Cross-validation is used to obtain a reliable estimate of the model accuracy for a given combination of features and model parameters.

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, prediction model, and number of top bands to use. 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.

This system proves to be robust for the following reasons. It is apparent that several of the considerations given at the beginning of Section 2.3 imply that there is an ordering from best to worst. The sampling reliability consideration suggests that some samples will be more representative than others. There is also an ordering in terms of noise, and one in terms of irrelevance. We do not know how many top bands to use, so we evaluate subsets containing varying numbers of top bands to find the optimal point, thus making the Hughes Phenomenon work for us, not against us. Finally, we do not know which ranking or prediction methods are best because of the NFL theorem, so we will test a variety of combinations. This allows us to implement

the system once, letting the computer do the work in any domains explored later. We therfore believe that our proposed system addresses all of the issues described in Section 2.3.

In this work, seven unsupervised ranking methods and three prediction methods are explored for both continuous and categorical prediction. 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 fuzzy k-nearest neighbors for categorical prediction. Linear regression, fuzzy k-nearest neighbors, and regression tree are used for continuous value prediction.

The following chapter presents detailed descriptions of both the unsupervised (3.1) and supervised (3.2, 3.3) methods used in this work. It should be stressed that the general framework does not depend on the specific methods employed. In fact, the orthogonality of the choice of methods is considered a major strength of ROWAS.

Chapter 3

Ranking and Prediction Methods

In this chapter we present the unsupervised ranking and supervised prediction methods we propose for use with ROWAS in a hyperspectral domain. The same ranking methods are used for both classification and regression experiments. As for predicition methods, linear regression and naïve bayes are unique to regression and classification, respectively. Standard implementations of k-nearest neighbors and decision trees are used for classification, and are then adapted for use in continuous value prediction (regression).

First, let us define our notation. Let $\mathbf{X}$ be an m dimensional vector of spectral intensity variables $X_1, X_2, ..., X_m$ related to a single pixel. $X_i = x_i = I(\phi, \lambda_i)$ is the measured intensity value of the band with central wavelength λ_i at geographic location ϕ. Let Y be a ground measurement known as the label. In a classification problem, $Y = c_j$ where $c_j \in C$ when C is the set of possible classes. In numeric prediction, Y takes on a particular continuous value y when sampled. A labeled data point, or *example*, is therefore an $(\mathbf{X}, Y)$ pair. A set S of examples is split into a subset of training examples T where $T \subset S$ and testing examples R where $R = S - T$ in order to measure accuracy. (T and R are randomly selected multiple times during cross-validation.)

In general, we relax the notation normally used to distinguish free variables from

samples in order to simplify the syntax. For example, $P(X|Y)$ is substituted for $P(X_1 = x_1, X_2 = x_2, ...X_m = x_m|Y = y)$.

3.1 Unsupervised Ranking Methods

3.1.1 Information Entropy

This method is based on evaluating each band separately using the information entropy measure ([26], chapter 3) defined below.

$$H(\lambda_i) = -\sum_{k=1}^{l} P(\nu_k^i) \ln P(\nu_k^i) \tag{3.1}$$

$$P(\nu_k^i) = P(\min_k^i \leq X_i < \max_k^i) \tag{3.2}$$

H is the entropy measure of the band with central wavelength λ_i. Equation 3.2 merely formalizes that the probability distribution function of the intensity value X_i is estimated via a histogram. Each bin k used to estimate the probability for a range of values of X_i is defined by $\{\min_k^i, \max_k^i\}$. l 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).

3.1.2 First Spectral Derivative

The bandwidth, or wavelength range, of each band is a variable in a hyperspectral sensor design [6, 5]. 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 Equation 3.3, where differences between

sampled adjacent bands X_i and X_{i+1} are summed for all examples in the set S. 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.

$$D_1(X_i) = \sum_S |x_i - x_{i+1}| \tag{3.3}$$

3.1.3 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 band X_i.

$$D_2(X_i) = \sum_S |x_{i-1} - 2x_i + x_{i+1}| \tag{3.4}$$

3.1.4 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.

$$\text{ContrastM}(X_i) = \sum_{j=1}^{m} |f_j - E(f)| * f_j \tag{3.5}$$

f is the histogram (estimated probability density function) of all contrast values computed across one band by using a Sobel edge detector ([26], Chapter 4). $E(f)$ is the sample mean of the histogram f. 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 $ContrastM$.

In the upcoming experimental results (Chapter 4), the contrast based unsupervised method utilized the fact that the hyperspectral examples extracted from the hyperspectral image were spatially ordered along a geo-spatial line (row). Our implemenation therefore assumed that the ordering of the examples could be used to determine adjacency for use in the Sobel edge detector.

3.1.5 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 ([27], 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 of band X_i based on the samples of set S.

$$\text{RatioM}(X_i) = \sum_S \left| \frac{x_i}{x_{i+1}} - E\left(\frac{X_i}{X_{i+1}}\right) \right| \tag{3.6}$$

3.1.6 Correlation Measure

One of the standard measures of band similarity is normalized correlation [20]. 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 [3]. 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 $CorM(X_i, X_j)$ represents the measure. E denotes an expected value and σ is a standard deviation.

$$\text{CorM}(X_i, X_j) = \frac{E(X_i * X_j) - E(X_i) * E(X_j)}{\sigma(X_i) * \sigma(X_j)} \tag{3.7}$$

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 [12], where spectrally adjacent blocks of correlated bands are represented in a selected subset.

3.1.7 Principal Component Analysis Ranking (PCAr)

Principal component analysis has been used very frequently for band selection in the past [27]. 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. These values are

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 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. The coefficients c_{ij} indicate the influence of the original dimension i on this 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 Equation 3.8.

$$PCAr(X_i) = \sum_j |\epsilon_j c_{ij}| \tag{3.8}$$

As the procedure for computing the eigenvalues and coefficients is both complex and available in most data analysis texts [20], it is omitted.

3.1.8 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.

3.2 Supervised Classification Methods

3.2.1 Naïve Bayes

Bayes law (3.9) provides the posterior probability of an event c_i occuring given that event $\mathbf{X}$ has occured based on the the prior probabilities of c_i and $\mathbf{X}$, as well as the posterior probability of event $\mathbf{X}$ given c_i. Here, this provides a means of calculating the probability of each possible class c_i given a spectral signature $\mathbf{X}$ and then selecting the class with the highest probability $P(c_i|\mathbf{X})$ as the prediction. $P(c_i)$ can easily be estimated from the set of training examples and $P(\mathbf{X})$, 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 $P(\mathbf{X}|c_i)$, conditional independence amongst attributes (here, spectral bands) is assumed (hence the name *Naïve* Bayes). This allows the use of Equation 3.10.

$$P(c_i|\mathbf{X}) = \frac{P(\mathbf{X}|c_i)P(c_i)}{P(\mathbf{X})} \tag{3.9}$$

$$P(\mathbf{X}|c_i) = \prod_k P(X_k|c_i) \tag{3.10}$$

In our implementation, the continuous variables X_k are binned, and estimated probabilities based on training data are stored in a histogram for every (c_i, X_k) pair for use in Equation 3.10. 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 2.3.1

3.2.2 Fuzzy K-Nearest Neighbors

K-nearest neighbors classifiers, sometimes called instance based classifiers [28], [14], 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

$$V(c_i) = \sum_{e \in \{e:Y_e = c_i\}} \frac{1}{d_e^w} \tag{3.11}$$

where Y_e is the class of training example e, and d_e is the euclidean distance

$$d_e = \sqrt{\sum_k (X_k - X_k^e)^2} \tag{3.12}$$

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 2.3.1.

3.2.3 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 [29].

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 $(n-1)$ 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 n 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:

$$H(T) = -\sum_{j} P(c_j|T) \ln P(c_j|T) \tag{3.13}$$

where the probability $P(c_j|T)$ is simply

$$P(c_j|T) = \frac{|\{e : e \in T, Y_e = c_j\}|}{|T|} \tag{3.14}$$

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

$$Gain(\nu) = H(T) - \sum_k \frac{|T_k|}{|T|} H(T_k) \tag{3.15}$$

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 2.3.1.

3.3 Supervised Continuous Prediction Methods

3.3.1 Linear Regression

The regression method used here is based on a multivariate linear regression [30, 14] that is used for predicting a single continuous variable Y given multiple continuous input variables $\{X_1, X_2, ...X_m\}$. The model building process can be described as follows. Given a set of training examples T, find the set of coefficients $\beta = \{\beta_0, \beta_1, ...\beta_m\}$ that gives the minimum value of $g(T)$, where

$$g(T) = \sum_{e \in T} (Y_e - Y'_e)^2 \tag{3.16}$$

Y_e is the observed output variable of a training example e and

$$Y'_e = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m \tag{3.17}$$

Y'_e is therefore the predicted value for Y_e given values for which, in this case, are reflectance values at varying wavelengths for the training example e. The problem as stated can be solved numerically using well-known matrix algebra techniques. Further details for finding are therefore omitted for the sake of brevity.

3.3.2 Fuzzy K-Nearest Neighbors (Continuous)

Like it's categorical counterpart, the continuous version of fuzzy k-nearest neighbors (kNN) works by first calculating the distance from the test example to all training examples (Equation 3.12). The target variables of of the k closest training examples are used to determine the final prediction. Unlike the categorical version, where each possible class has its own "vote" weighted by the distances, there is a single summation of the target values weighted by the inverse distance (Equation 3.18).

$$Y'_e = \sum_{i \in \text{NearestNeighbors}}^{k} \frac{Y_i}{d_e^w} \tag{3.18}$$

Once again, the exponent weight w and the number of nearest neighbors used to calculate the prediction k are the control parameters to be optimized.

3.3.3 Regression Tree

A regression tree has the same fundamental structure as the C4.5 decision tree presented in Section 3.2.3, but is modified to handle a continuous target variable.

The first modification is the use of variance (Equation 3.19) instead of entropy of the target variable to evaluate the improvement gained from a split. Like the categorical version, the values of the discrimination metric for the example subsets created from a split are summed, weighted by the number of examples in each subset (Equation 3.20). The potential split with the greatest improvement in the variance summation is selected and the algorithm repeats on the child nodes.

$$V(T) = \sum_{e \in T} \frac{(Y_e - E(Y_e))^2}{|T| - 1} \tag{3.19}$$

$$Gain(S) = V(T) - \sum_{k} \frac{|T_k|}{|T|} V(T_k) \tag{3.20}$$

The tree building process halts when the minimum permitted gain, τ_i, is not satisfied by any potential splits, or the minimum number of examples per node, τ_e, has been met.

When a node has been marked as a leaf, it requires a mechanism to make a continuous valued prediction. There are two possibilities. The first is a linear regression model (from Section 3.3.1) built using the training examples that have evaluated to that leaf. A boolean control parameter ρ determines which features are used in the regression model. If $\rho =$ true, then all of the features presented to the regression tree are used. If $\rho =$ false, then only those features that were used as split tests to reach that leaf are used. If the linear regression should fail because of insufficient training examples in the leaf or because it simply did not converge, then a mean model is used as the continuous value prediction mechanism. A mean model is simply a model that returns the average target value of a training set, ignoring all input features. It is therefore quite naïve, but can return a *reasonable* prediction given a very small training set.

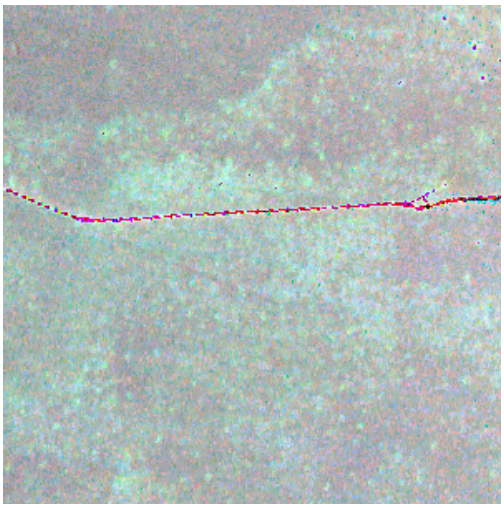
Chapter 4

Experimental Results

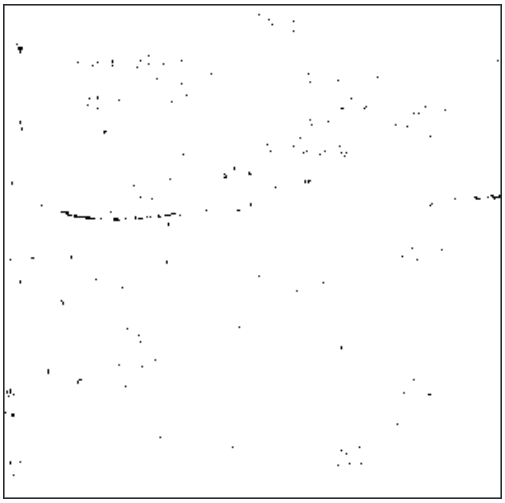
We now present an empirical study of the performance of ROWAS in one classification task and one regression task. Sections of this chapter are reproduced from [31] and [32].

4.1 Classification of AVIRIS Data

We obtained a data set that consisted of spectral measurements from an AVIRIS [33] sensor and manually collected labels of the grass type of scanned regions [34]. 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 consisted of {Unclassified, Black Grama, Blue Grama, Road, Black Grama/Green Veg Mixed, Blue Grama/Green Veg Mixed}. Figure 4.1 shows both a spectral sample of the field and the locations of the ground labels.



(a) Color composite of AVIRIS bands 400 nm, 1300 nm, and 2209 nm.



(b) Binary image of valid labels. Black pixels are the locations that have corresponding ground measurements.

Figure 4.1: Visualizations of the AVIRIS hyperspectral data and grass labels. Taken October 20, 1999

4.1.1 AVIRIS Results

The top score for each supervised, unsupervised method pair is given in Table 4.1. The score is the sample mean fraction of misclassifications obtained from the final 12-fold cross-validation performed for every set of top ranked features. It is therefore effectively the average percent error. 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. Only the first 100 bands were tested as the computational expense became too severe at that point. Each model optimization was allowed 100 random parameter sets.

The graphs of Figures 4.2 - 4.4 show the complete results for the three unsupervised methods with the best scores for each supervised method. Also included are random rankings 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

Table 4.1: The number (count) of top ranked bands used to achieve the best average fraction of misclassifications, and the error itself.

	Naïve Bayes		K-Nearest Neighbors		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	.027	76	.116	92

light blue plots correspond to these trials, and a bold red line corresponds to their average. The best random ranking and the average were considered the baselines for comparison.

Naïve Bayes

Naïve Bayes (Figure 4.2) 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 Figure 4.2 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

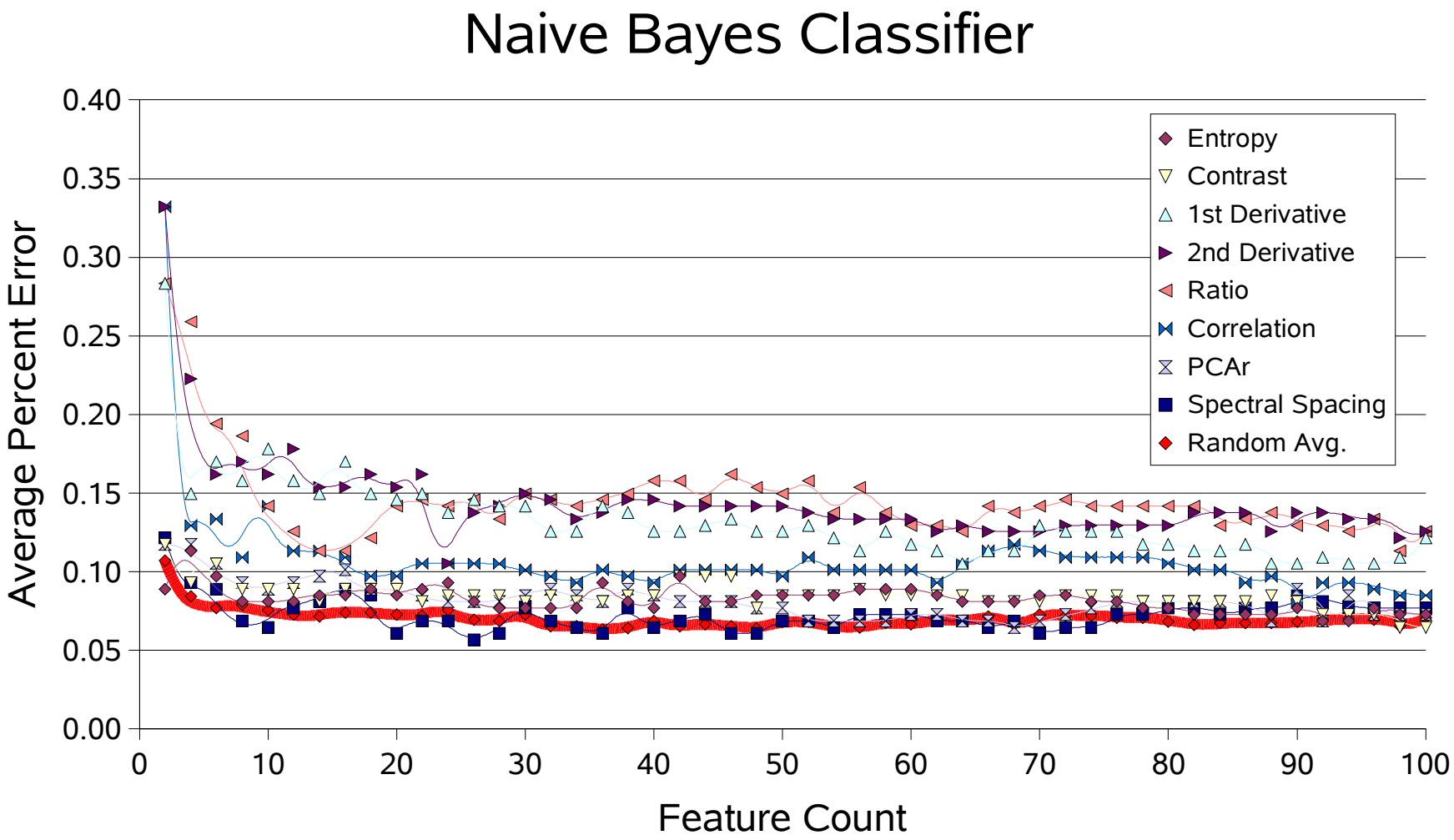


Figure 4.2: Naïve bayes results for classification of the AVIRIS data using intelligent ranking methods. Each plot represents the trend of a sequentially growing number of top bands as determined by a particular ranking method.

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.

Fuzzy K-Nearest Neighbors

Next was the instance based classifier, with the best results shown in Figure 4.3. Instance based classifiers can be finely tuned to a data set due to its parameters that can vary the behaviour 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

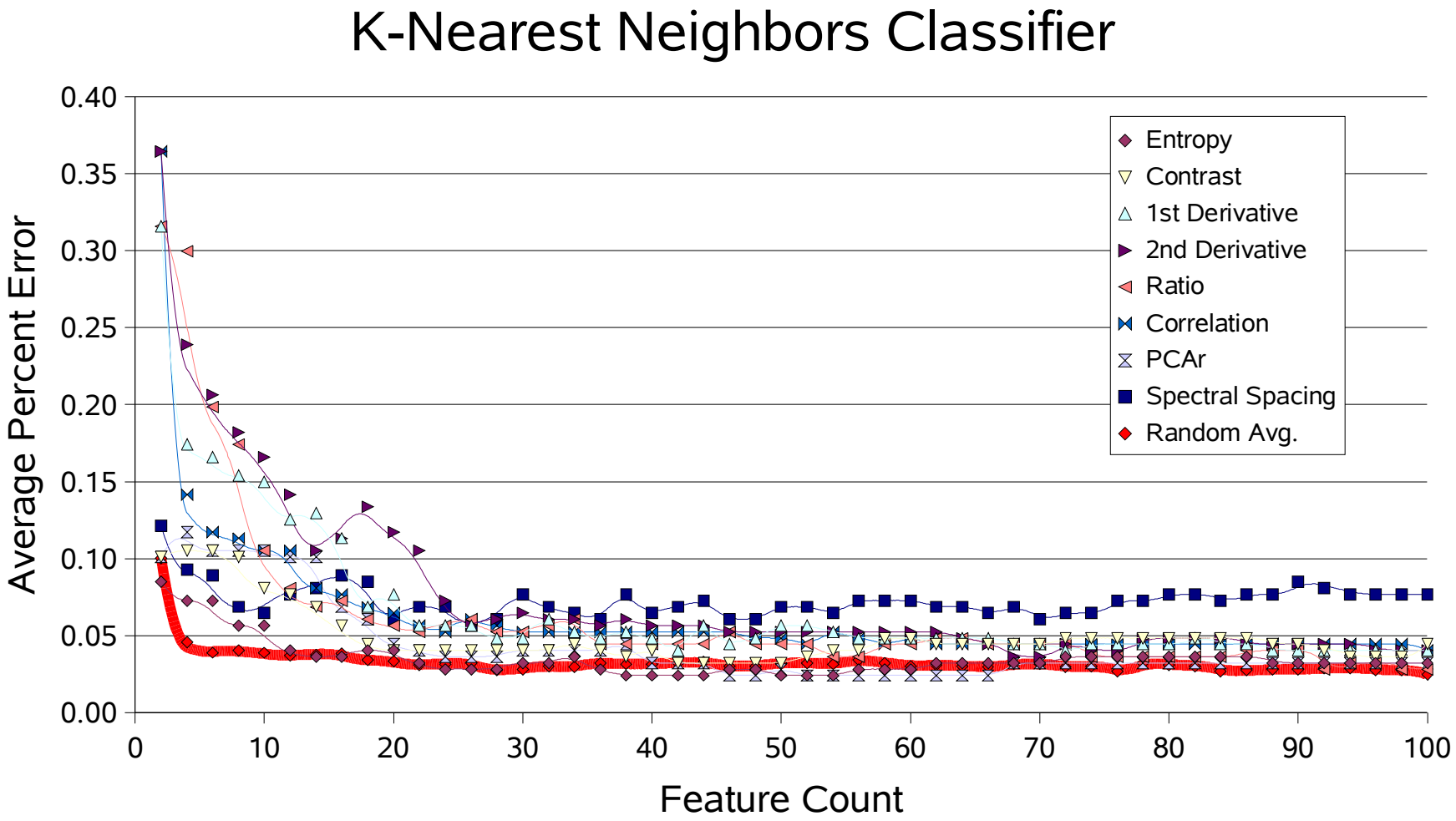


Figure 4.3: K-nearest neighbors results for classification of the AVIRIS data using intelligent ranking methods. Each plot represents the trend of a sequentially growing number of top bands as determined by a particular ranking method.

the information content of the bands. They do not take into account the redundancy of bands, and are demonstrated to be 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 model using the spectral 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 (4.1.2).

C4.5 Decision Tree

Finally, the C4.5 decision tree results are given in Figure 4.4. 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

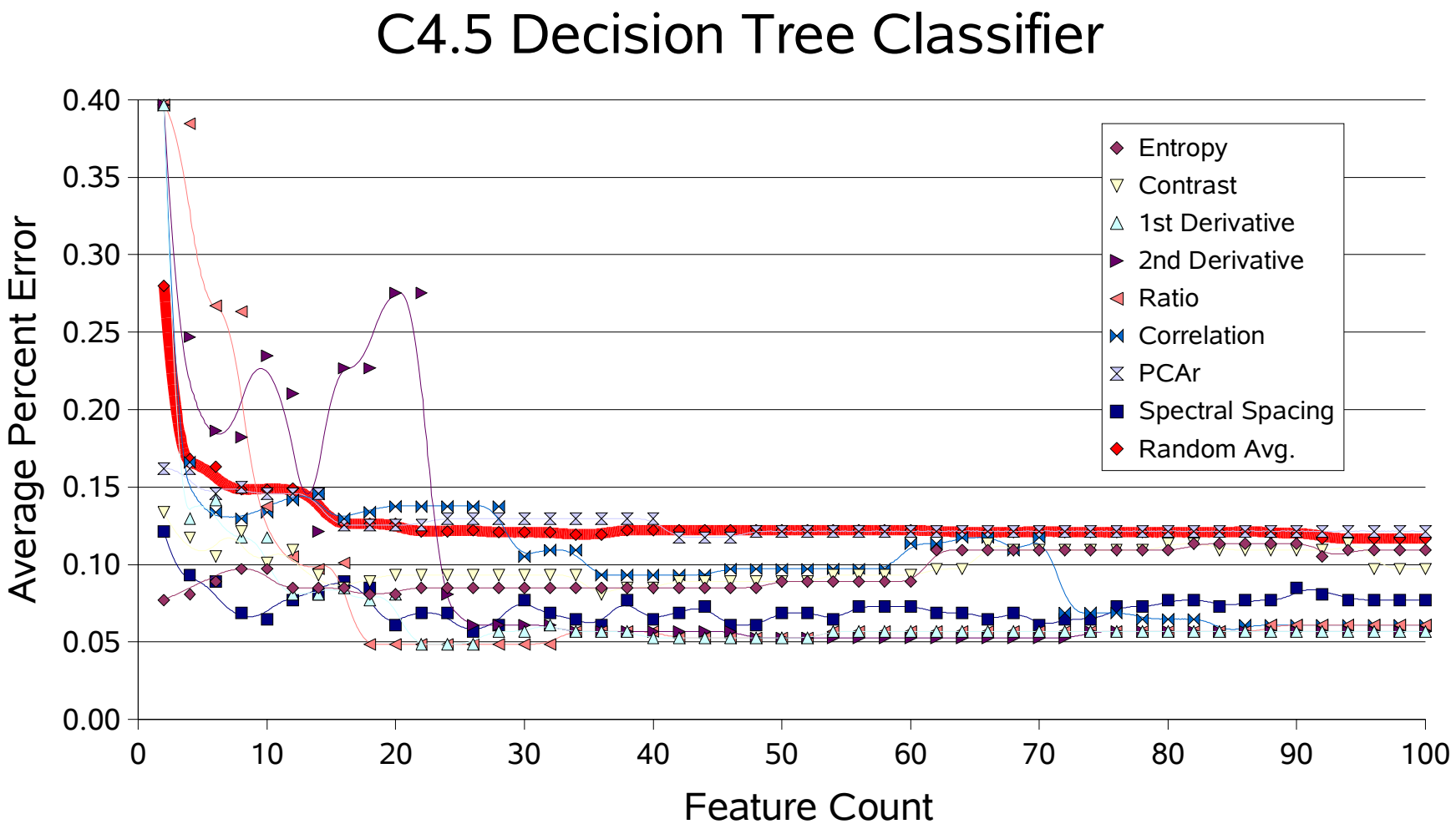


Figure 4.4: Decision tree results for classification of the AVIRIS data using intelligent ranking methods. Each plot represents the trend of a sequentially growing number of top bands as determined by a particular ranking method.

Table 4.2: 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

features.

4.1.2 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 [35] 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 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 (4.1).

$$P(\geq x) = \sum_{r=x}^{m} \binom{m}{r} (\pi_b)^r (\pi_a)^{(m-r)} \tag{4.1}$$

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 (4.1) 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 (4.1) 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 4.2. 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 4.2 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.

4.2 Regression of RDACS Data

The hyperspectral image data used in this section were collected from an aerial platform with a Regional Data Assembly Centers Sensor (RDACS), model hyperspectral (H-3), which is a 120-channel prism-grading, push-broom sensor developed by NASA. Each image has 2500 rows, 640 columns, and 120 bands per pixel. The 120 bands correspond to the visible and infrared range of 471 to 828 nm, recorded at a spectral resolution of 3 nm. The motivation for choosing the wavelength range came from the agricultural application domain where the 400-900 nm wavelength range responds to plant characteristics very well [36] (Chapters 2-2 and 5-2) and has been used for vegetation sensing in the past [9]. By selecting this wavelength range, the data analysis avoids issues related to water absorption bands (1400 nm and 1900 nm). In the particular range we compensated only for low reflectance in the blue (450 nm) and red (650 nm) wavelength sub-ranges due to the two chlorophyll absorption bands [27] (Chapter 17.4) during reflectance calibration. While our experiments dealt with

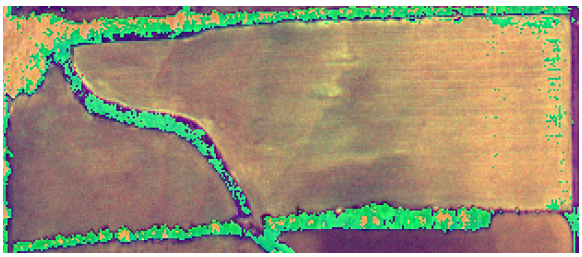


Figure 4.5: An RGB approximation of the gvillo field data taken April 26, 2000.

images of bare soil, we used a sensor that is optimal for vegetation observation as that is what is likely to be available in agricultural applications (for the reasons given above). Indeed, the experimental data set used in this study came from a series of images taken over the entire growing season that were collected to study the relationship between hyperspectral information and both bare soil properties before crops had emerged and crop properties when they were present. For application specific interpretations of data, each band index of the hyperspectral image was converted to the band central wavelength by applying the following formula:

The images were collected from altitudes in the range of 1200 m to 4000 m on April 26, 2000. The spatial resolution of the images is approximately 1-m for the processed Gvillo field located near the city of Columbia in the central part of Missouri. The images were pre-processed to correct for geometrical distortions, calibrated for sensor noise and illumination, and geo-registered [36] (Chapter 2-7). However, the images were not pre-processed for any atmospheric corrections [27] (Chapter 10-4). An image of the Gvillo site is shown in Figure 4.5.

Ground measurements of several variables (e.g., conductivity, elevation, organic matter, phosphorous content) were collected by the Illinois Laboratory Agricultural Remote Sensing (ILARS) using the Veris profiler 3000 made by Veris Technologies, Salina, KS, and the data were provided by Dr. Tian. The hyperspectral images provided by Spectral Visions, a non-profit research organization funded by the NASA Commercial Remote Sensing Program, were geo-registered with the ground measurements by Dr. Gopalapillai (Department of Biological and Agricultural Engineering,

University of Arkansas) and both ground and aerial measurements formed a training data set covering about 19,000 m2 of the Gvillo field. We used the training data with 190 examples from the hyperspectral imagery collected at 4000 m altitude for evaluating the band selection methods. The training data contained these hyperspectral values and associated ground values of soil electrical conductivity. The field coverage on the date of data collection was bare soil.

Among all ground variables, we anticipated to find relationships between hyperspectral values (reflected part of the electro-magnetic (EM) waves in the wavelength range [471nm, 828 nm]) and surface/field characteristics that change electric and magnetic properties according to the EM theory of wave propagation [37] (Chapter 5). Thus, electrical conductivity appeared as the number one candidate among other variables. We verified with a simple linear correlation method that there exists a significant enough correlation (around 0.5) between the conductivity variable and hyperspectral values (190 conductivity values were correlated with 190 hyperspectral values for each band to obtain 120 correlation values averaging near 0.5). The conductivity values ranged from [22.4262, 52.66] miliSiemens per meter (mS/m) with the sample mean equal to 36.10836 and the standard deviation equal to 5.212215. Based on the known classification of soil properties [38] as a function of conductivity with approximate class conductivity ranges of sand (0,2], silt [2, 20] and clay [10, 1000], we concluded that the ground soil consisted of silt and clay soil types. Soil electrical conductivity is an important characteristic considered for crop yield prediction in the agricultural application. Electrical conductivity indirectly characterizes several important soil characteristics including soil texture (the relative amount of sand-silt-clay) and salinity, which affects the crops ability to acquire water.

Table 4.3: The number (count) of top ranked bands used to achieve the best sample mean absolute error, and the error itself in milliSiemens/meter (mS/m).

	Linear Regression		K-Nearest Neighbors		Regression Tree	
	Error	Count	Error	Count	Error	Count
Entropy	1.99	6	2.71	112	1.90	6
1^{st} Deriv.	1.99	8	2.65	14	1.98	8
2^{nd} Deriv.	2.01	8	2.66	106	1.96	6
Contrast	2.08	14	2.70	110	2.02	14
Ratio	2.05	10	2.68	108	1.92	6
Correlation	2.03	10	2.48	16	1.90	8
PCAr	2.10	16	2.71	108	2.07	16
Spectral Spacing	2.04	8	2.55	4	1.93	6
Best Random	2.01	6	2.53	8	1.89	6
Average Random	2.04	12	2.67	40	1.97	6

4.2.1 RDACS Results

As before, the top score for each supervised, unsupervised method pair is given in Table 4.3. 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 experiment was allowed to run until all 120 bands had been evaluated. Each model optimization was allowed 300 random parameter sets to try.

Linear Regression

Nearly all of the ranking methods are competitive when matched with a linear regression model (Figure 4.6). The top methods are entropy and 1^{st} derivative, but the

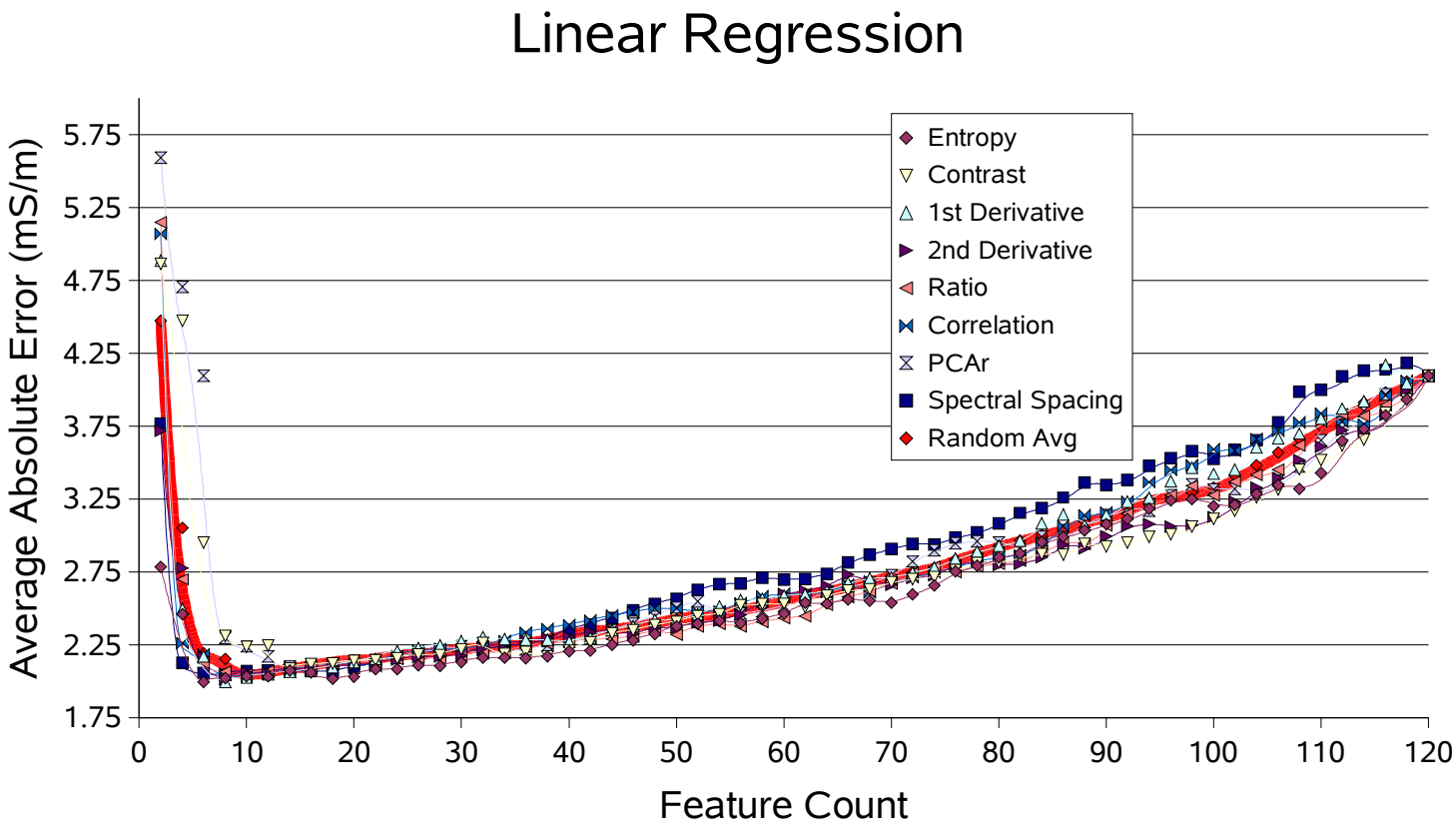


Figure 4.6: Linear regression results for regression of the RDACS data using intelligent ranking methods. Each plot represents the trend of a sequentially growing number of top bands as determined by a particular ranking method.

difference in average error between those methods and, say, 2^{nd} derivative or the top random method is negligible. As compared to the other supervised methods, linear regression performs nearly as well as the regression tree and significantly better than the nearest neighbors method.

Also of interest is that linear regression shows the most pronounced parabolic behavior. The optimal number of bands is reached quickly and then steadily declines. We believe this shows that the majority of useful information in the hyperspectral imagery can be gleened from just a few bands. It also reflects a trend we expected when comparing the classification to regression methods. The classification methods are more resilient to small deviations in their predictions. This is because classification methods can be thought of as predicting a continuous valued probability for each class, and returning the best. A small change in these probabilities will not cause a change in accuracy, so long as the change is not so extreme that the probabilities change their ordering from highest to lowest. In contrast, a set of small detrimental changes to a set of regression predictions will each add to the final reported error shown in the graphs. This phenomenon is also apparent when comparing the C4.5 Decision Tree and the Regression Tree.

Fuzzy K-Nearest Neighbors

The fuzzy k-nearest neighbors did the least well on the continuous value prediction problem. Not only was the overall error high, the optimal number of bands using many of the ranking methods was quite high due to the asymptotic nature of the error function (Figure 4.7). The asymptotic behavior was likely due to the local search nature of the algorithm. Because only training examples near a test example in the feature space are considered for a prediction, the higher dimensionality would not cause problems if the lower ranked features contained the same information as the higher ranked features (which the analysis of linear regression seems to imply).

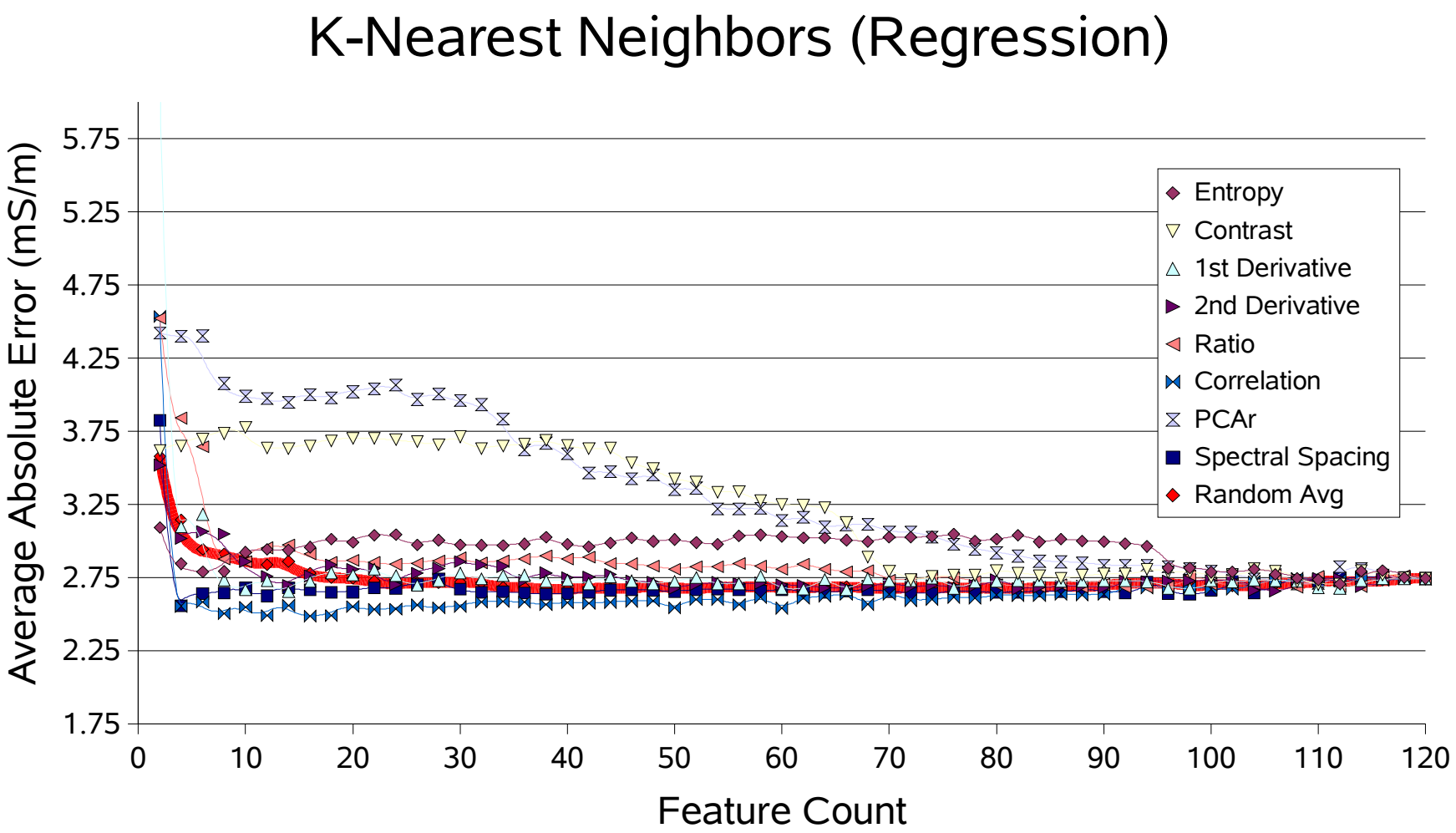


Figure 4.7: Fuzzy k-nearest neighbors results for regression of the RDACS data using intelligent ranking methods. Each plot represents the trend of a sequentially growing number of top bands as determined by a particular ranking method.

In fact, the information contained in the lower ranked bands seems to smooth out the noise contained in the better ranked bands, causing a slight increase in accuracy as the number of top ranked bands used in prediction is increased. This was not a very significant decrease in error, however, as the approximate error gradient for curves when the band count was larger than 6 was less than 0.002.

The few methods that were able to overcome the aymptotic, but poor, performance were correlation, spectral spacing, and one of the random rankings. The two intelligent methods of these work by selecting the least similar bands, and is likely the random ranking that performed well coincidently did the same. We can hypothesis that the optimal number of bands determined by the Hughes Phenomenon was quite low (under 14), and only these best methods were able to present all of the relevant information to the learning algorithm in 14 bands or less.

Regression Tree

Our final supervised prediction method was the regression tree, shown in Figure 4.8. It provided the best regression accuracy with some unusual trends. The entropy, correlation, and best random ranking methods performed the best with errors near 1.90 mS/m using either 6 or 8 bands. That fact that entropy is an information content based method and correlation is a redundancy measure shows that several methods may exist that provide good empirical performance in any given domain.

The plots in Figure 4.8 quickly reach an optimal point between 6 and 16 bands, then degrade continuously before plateauing near 80 bands. The form of the graph before the plateau is quite similar to that of the linear regression graph of Figure 4.6. This is not surprising as the final predictions made by the regression tree are simply linear regression models that use certain subsets of features and training examples as determined by the tree structure. The plateaus are easily explained if we recall the behavior of the regression tree when a leaf has too few examples relative to the

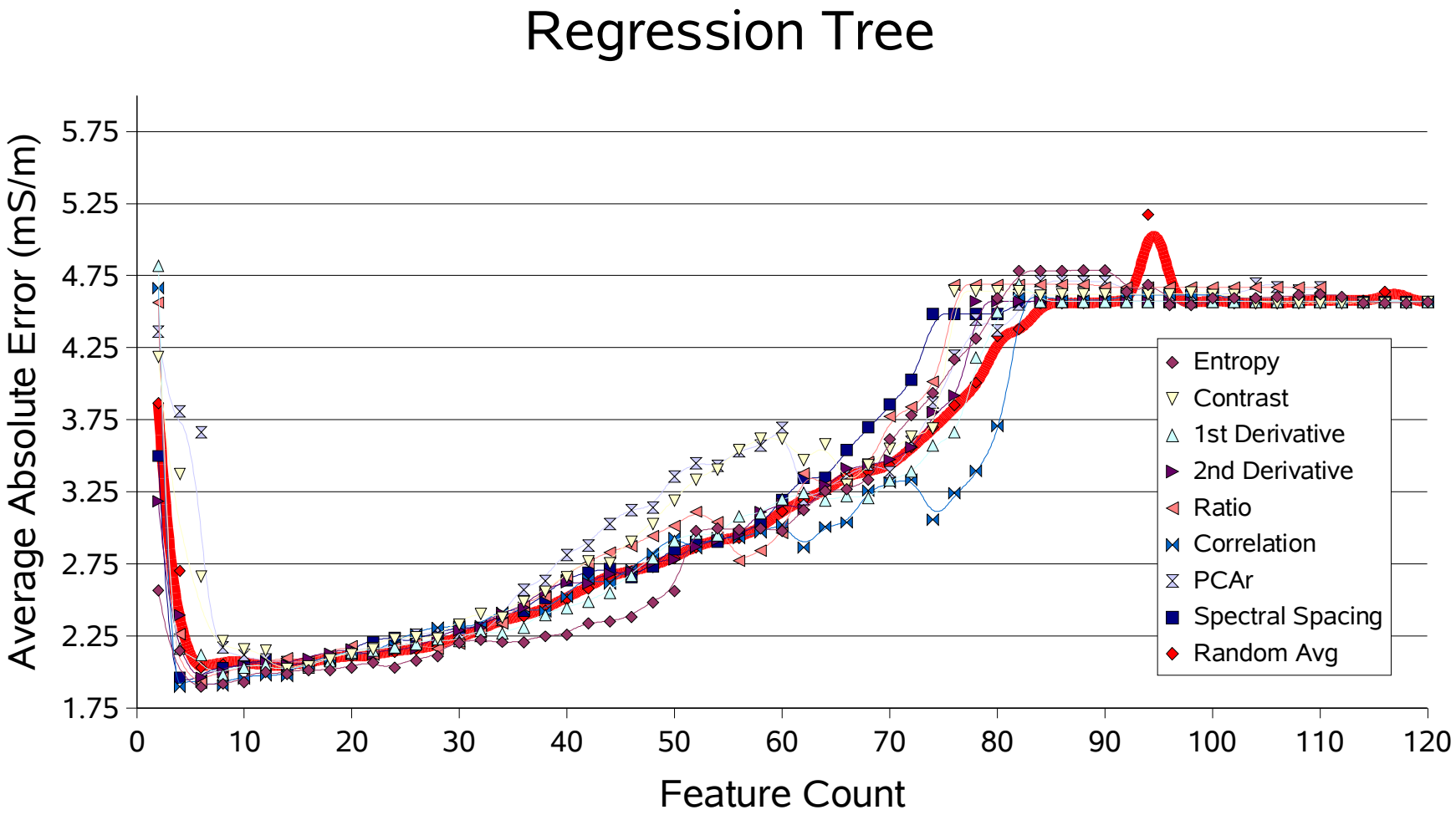


Figure 4.8: Regression Tree results for regression of the RDACS data using intelligent ranking methods. Each plot represents the trend of a sequentially growing number of top bands as determined by a particular ranking method.

dimensionality of the data set and the linear regression fails. In such a situation a mean model is returned, which will yield a single accuracy score if the data set remains constant. A single split of the 176 example training set used in 12-fold cross-validation would lead to leaves containing about 90 examples. It is not surprising, therefore, that a linear regression model built in 80 dimensions would not converge with so few examples relative to the number of dimensions, and the mean model would be returned. As is shown in the graph, this is independent of the ranking method.

Chapter 5

Summary and Conclusions

In this thesis we have (a) identified the constraints a modern feature selection system must address, (b) developed a methodology that works within those constraints, and (c) empirically tested the methodology in the domain of hyperspectral image analysis.

5.1 Feature Selection Considerations

The issues facing a feature selection strategy are relevance, redundancy, sample quality, computation time, and user effort. These issues were presented in detail in Chapter 2. The considerations can be summarized as follows.

Noise is an obvious problem as it obscures the true nature of the data and their internal relationships. Noise can be in the form of sensor noise, but irrelevant features can also be considered noise as they increase the complexity of the problem without adding any useful information (Section 2.1.1) . The Hughes Phenomenon also affects the selection of best features. It states that an increase in dimensions without an increase in data points reduces the validity of density measurements, and therefore can degrade the accuracy of prediction models that use them. The result is that there is an optimal number of features to use for a particular sized data set (in terms of sample size) with a given modeling technique (Section 2.1.2). Furthermore, the central

limit theorem tells us that the parameters of a sample's distribution will vary from the parameters of their true populations according to known rules. We can therefore expect that as the number of available features to choose from goes up, we can expect the likelihood of having as many "good" features as the Hughes Phenomenon dictates to also increase (Section 2.1.1).

Unfortunately, determining the best bands can not be determined directly, as the true population parameters are unknown. Because it is therefore necessary to use some form of heuristic search for the best bands, the NFL theorem comes into play. The NFL theorem tells us that no optimization technique is superior in all domains. When applied to the problem at hand, it states that the best feature selection technique cannot be known in advance, unless some reliable domain knowledge exists, which we assume is not the case in this thesis (Section 2.2). We purposely make this assumption for the following reason. Domain experts' time is vastly more valuable than computer resources, and the gap will only widen as mass-produced computation systems continue to fall in price. If the only reason to tap into a domain expert's time to save the computational burden of trying several different feature selection techniques, we believe the costs associated with doing so are unjustified (Section 2.2).

5.2 ROWAS

To overcome the considerations given above, we proposed a general framework for a joint feature subset and prediction model system. We have called this methodology Rank Ordered with Accuracy Selection (ROWAS), which was presented in Section 2.3.1. The method matches a set of computationally efficient, unsupervised ranking methods with a set of supervised prediction methods and a somewhat expensive, but highly accurate, error evaluation method. As sets of an incrementally larger number of top ranked bands are given to the supervised methods and evaluated, the optimal

number of bands for a given (ranking, prediction method) pair emerges. The overall best combination of feature subset and prediction method has a high chance of being near the theoretical optimal accuracy for a given data set. We can make this claim if a sufficient number of (ranking, prediction method) pairs are explored to satisfy the restrictions of the No Free Lunch Theorem.

5.3 Results

We tested our methodology on two data sets from the hyperspectral image analysis domain. The classification problem used AVIRIS data to predict the grass type labels in an image taken in New Mexico. We also used our methodology in a continuous value prediction problem in the agriculture domain. Data from an RDACS sensor was used to predict the electrical conductivity of soil in an early season production field.

In the classification problem, the combination of the spectral spacing ranking method and k-nearest neighbors classifier performed the best, although one of the random rankings performed slightly better (also with kNN). A statistical significance test, however, showed that the small differences in error among the top methods were not conclusively significant. The test was even rather conservative, as it took into account only those examples that were classified differently, and was therefore independent of the total number of examples in the test set.

For the regression problem, the regression tree coupled with one of the random rankings was marginally better than when coupled with the entropy or correlation rankings. While no formal statistical analysis of was done in this part of the study, the average difference between the random and next best rankings was .01 mS/m in predicting a variable that ranged from 22.42 mS/m to 52.66 mS/m. We do not believe these difference are significant, and consider the three top ranking methods

to be equally valid when used with the regression tree.

5.4 Conclusions

We believe the prediction accuracies obtained are quite good and would meet any real world performance requirements. The fact that the same ranking methods were able to be used without modification in two distinct hyperspectral prediction problems also reflects well on the methodology. We believe such code reuse allows the end user experience to take on a simplicity that is a necessity for any successful, real world application of a machine learning technique. By removing the burden of feature subset and prediction method selection from the end user, and placing that burden on increasingly inexpensive computing technology, the application of machine learning can be moved from corporate and university research departments into the hands of those concerned about the subject being analyzed, not cutting edge machine learning technology. We believe our methodology can be the basis for applications that fulfill this key requirement of successful technology.

Appendix A

Results of Random Rankings

This section provides feature count versus error plots when random rankings were used in the experiments as described in Chapter **??**. First, the three supervised classification methods of Section 3.2 are used with random rankings of the AVIRIS data of Section 4.1. Second, the three supervised regression methods of Section 3.3 are used with random rankings of the data of Section 4.2.

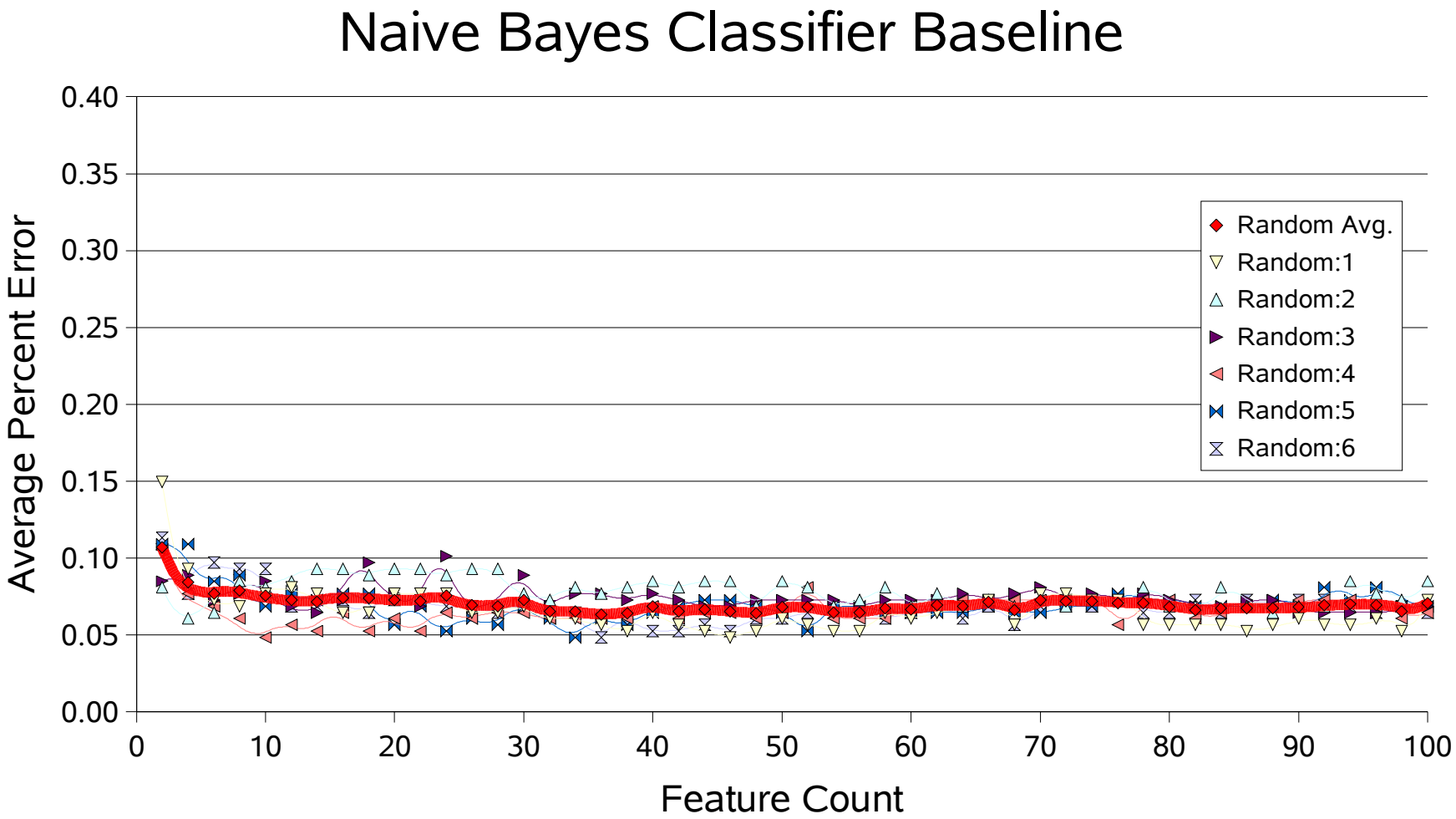


Figure A.1: Naïve bayes results for classification of the AVIRIS data using intelligent ranking methods. Each plot represents the trend of a sequentially growing number of top bands as determined by a particular ranking method. The bold red Random Average plot is identical to the Random Average plot in Figure 4.2

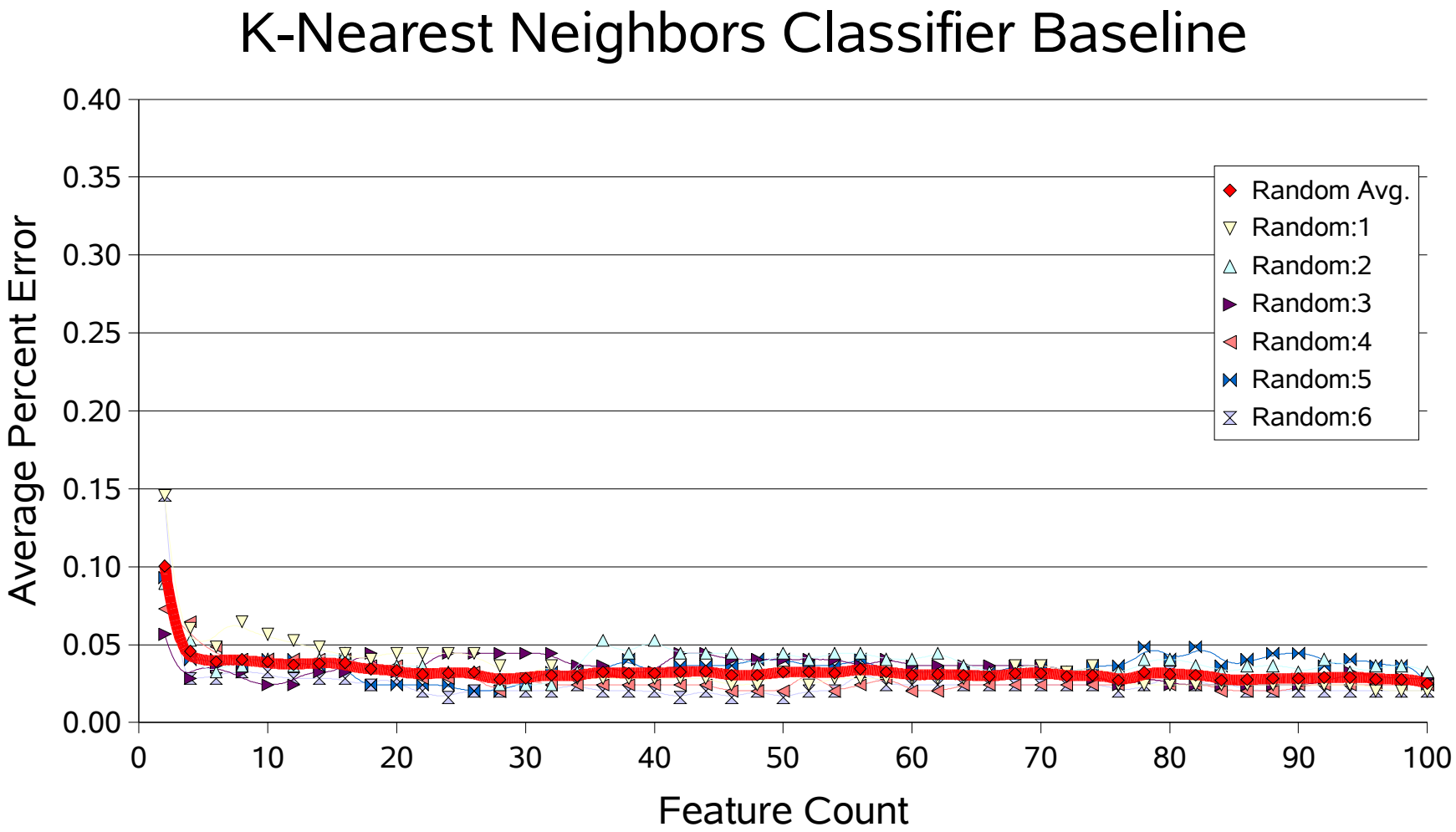


Figure A.2: K-nearest neighbors results for classification of the AVIRIS data using intelligent ranking methods. Each plot represents the trend of a sequentially growing number of top bands as determined by a particular ranking method. The bold red Random Average plot is identical to the Random Average plot in Figure 4.3.

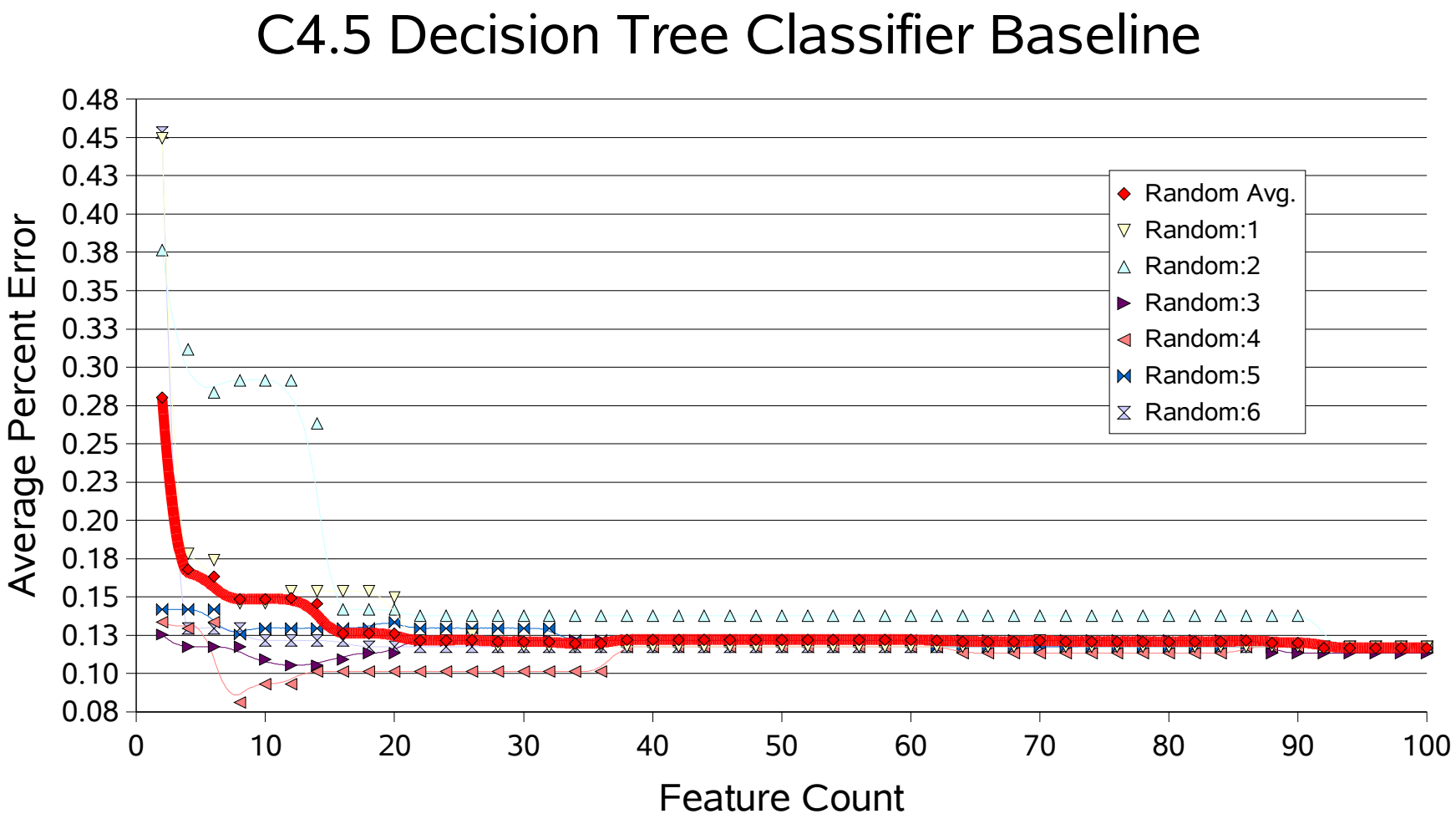


Figure A.3: Decision tree results for classification of the AVIRIS data using random ranking methods. Each plot represents the trend of a sequentially growing number of top bands as determined by a particular ranking method. The bold red Random Average plot is identical to the Random Average plot in Figure 4.4.

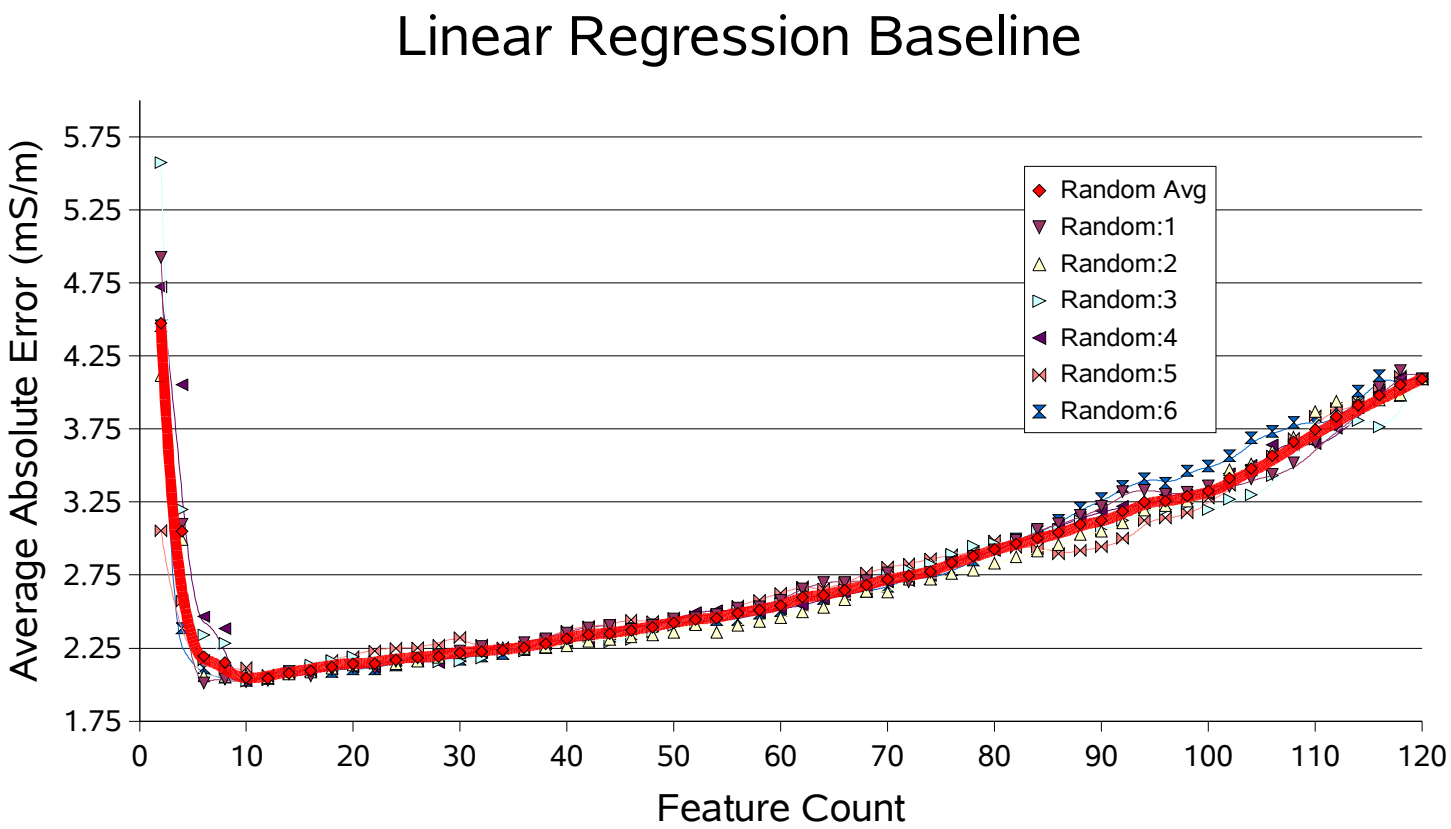


Figure A.4: Linear regression results for regression of the RDACS data using random ranking methods. Each plot represents the trend of a sequentially growing number of top bands as determined by a particular ranking method. The bold red Random Average plot is identical to the Random Average plot in Figure 4.6.

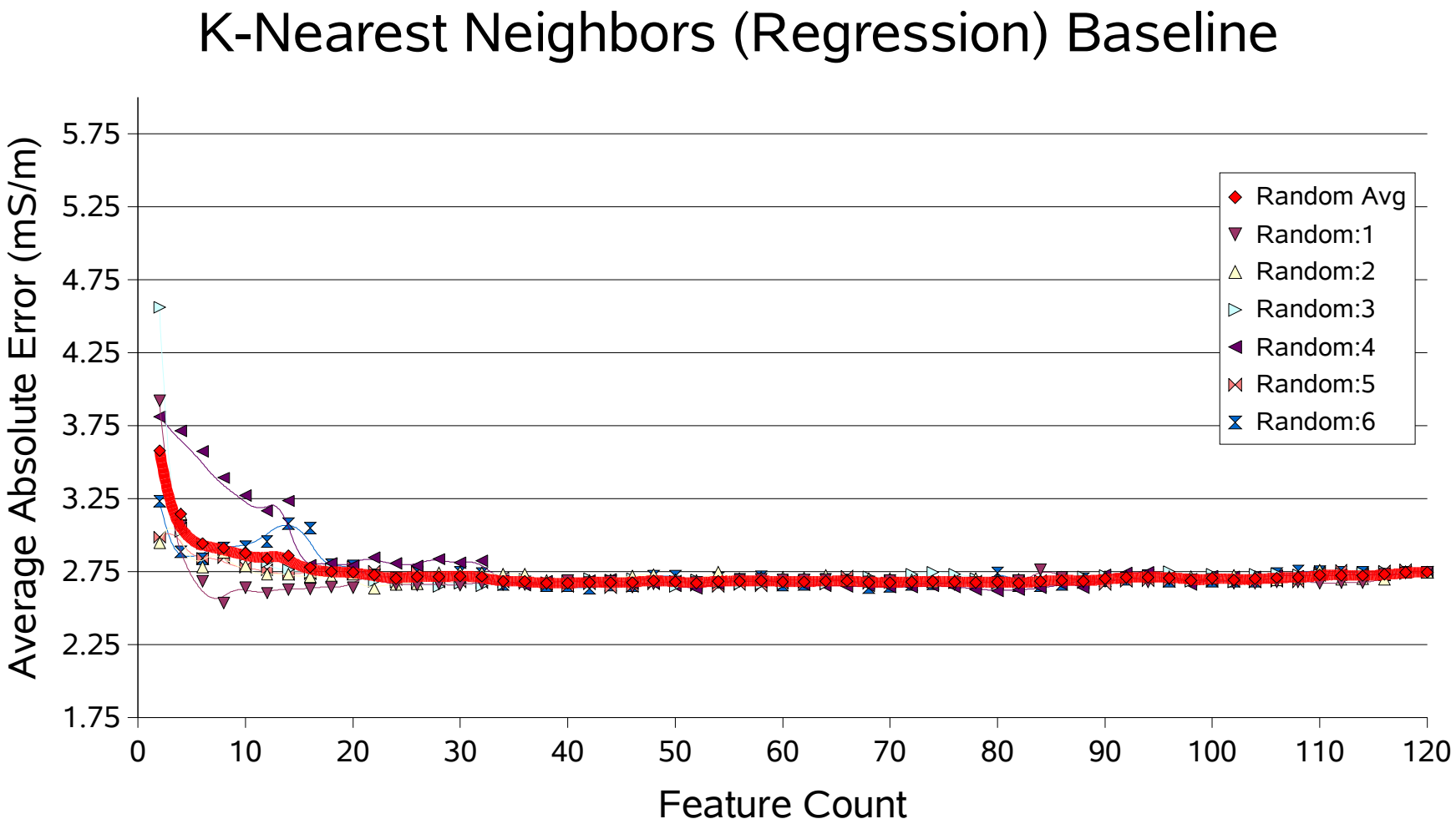


Figure A.5: Fuzzy k-nearest neighbors results for regression of the RDACS data using random ranking methods. Each plot represents the trend of a sequentially growing number of top bands as determined by a particular ranking method.The bold red Random Average plot is identical to the Random Average plot in Figure 4.7.

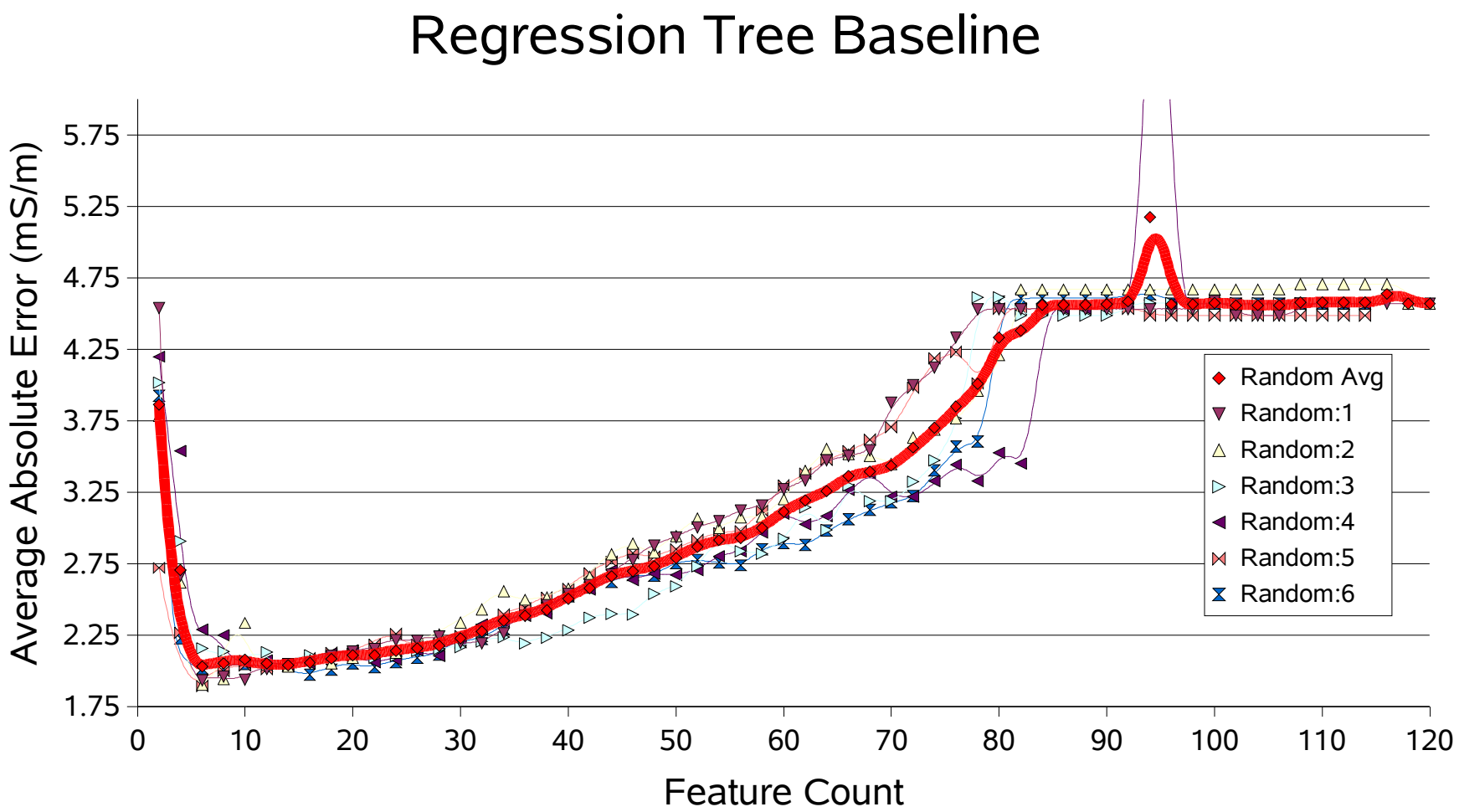


Figure A.6: Regression Tree results for regression of the RDACS data using random ranking methods. Each plot represents the trend of a sequentially growing number of top bands as determined by a particular ranking method.The bold red Random Average plot is identical to the Random Average plot in Figure 4.8

References

[1] F. Csillag, L. Pasztor, and L. L. Biehl. "Spectral band selection for the characterization of salinity status of soils." *Remote Sensing of Environment* **43** 231–242 (1993).

[2] Y. Yamagata. "Unmixing with subspace method and application to hyper spectral image." *Journal of Japanese Society of Photogrammetry and Remote Sensing* **35** 34–42 (1996).

[3] T. Warner, K. Steinmaus, and H. Foote. "An evaluation of spatial autocorrelation-based feature selection." *International Journal of Remote Sensing* **20** 1601–1616 (1999).

[4] E. Merényi, W. H. Farrand, L. E. Stevens, T. S. Melis, and K. Chhibber. "Mapping colorado river ecosystem resources in glen canyon: Analysis of hyperspectral low-altitude aviris imagery." In *Proceedings of ERIM, 14th International Conference and Workshops on Applied Geologic Remote Sensing*, pp. 44–51. Las Vegas, NV (November 2000).

[5] D. J. Wiersma and D. A. Landgrebe. "Analytical design of multispectral sensors." *IEEE Trans. Geosci. Remote Sensing* **18** 180–189 (1980).

[6] J. C. Price. "Band selection procedure for multispectral scanners." *Applied Optics* **33** 3281–3288 (1994).

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

[8] E. Merényi, R. B. Singer, and J. S. Miller. "Mapping of spectral variations on the surface of mars from high spectral resolution telescopic images." *Icarus, International Journal of Solar System Studies* pp. 280–295 (1996).

[9] S. Gopalapillai and L. Tian. "In-field variability detection and yield prediction in corn using digital aerial imaging." *Transactions of the ASAE* **42** 1911–1920 (1999).

[10] R. Pu and P. Gong. "Band selection from hyperspectral data for conifer species identification." In *Proceedings of Geoinformatics'00 Conference*, pp. 139–146. Monterey Bay, CA (June 2000).

[11] G. Healey and D. A. Slater. "Invariant recognition in hyperspectral images." In *IEEE Proceedings of CVPR99*, pp. 438–443 (1999).

[12] X. Jia and J. A. Richards. "Efficient maximum likelihood classification for imaging spectrometer data sets." *IEEE Trans. Geosci. Remote Sensing* **32** 274–281 (March 1994).

[13] P. J. Withagen, E. den Breejen, E. M. Franken, A. N. de Jong, and H. Winkel. "Band selection from a hyperspectral data-cube for a real-time multispectral 3ccd camera." In *Proceedings of SPIE AeroSense, Algorithms for Multi-, Hyper, and Ultraspectral Imagery VII*, volume 4381. Orlando, FL (April 2001).

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

[15] Q. Z. Jackson and D. Landgrebe. *Design Of An Adaptive Classification Procedure For The Analysis Of High-Dimensional Data With Limited Training Sample.* Ph.D. thesis, Purdue University (2001).

[16] R. Kohavi and G. H. John. "Wrappers for feature subset selection." *Artificial Intelligence* **97** 273–324 (1997).

[17] G. H. John, R. Kohavi, and K. Pfleger. "Irrelevant features and the subset selection problem." In *International Conference on Machine Learning*, pp. 121–129 (1994).

[18] J. Yang and V. Honavar. "Feature subset selection using A genetic algorithm." In *Genetic Programming 1997: Proceedings of the Second Annual Conference*, J. R. Koza, K. Deb, M. Dorigo, D. B. Fogel, M. Garzon, H. Iba, and R. L. Riolo, editors, p. 380 (Morgan Kaufmann, Stanford University, CA, USA, 1997).

[19] I. Tsamardinos and C. Aliferis. "Towards principled feature selection: Relevancy, filters, and wrappers." (January 2003).

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

[21] B.-C. Kuo and D. A. Landgrebe. "A robust classification procedure based on mixture classifier and nonparametric weighted feature extraction." *IEEE Trans. Geosci. Remote Sensing* **40** 2486–2494 (November 2002).

[22] H. Du, H. Qi, X. Wang, R. Ramanath, and W. E. Snyder. "Band selection using independent component analysis for hyperspectral image processing." In *Proceedings of the AIPR Workshop* (October 2003).

[23] G. F. Hughes. "On the mean accuracy of statistical pattern recognizers." *IEEE Trans. Inform. Theory* **IT-14** (January 1968).

[24] S. K. Murthy. *On Growing Better Decision Trees from Data.* Ph.D. thesis.

[25] W. Mendenhall, R. L. Scheaffer, and D. D. Wackerly. *Mathematical Statistics with Applications*, 3rd edition (Morgan Kaufmann Publishers, San Francisco, 1993).

[26] J. C. Russ. *The Image Processing Handbook*, 3rd edition (CRC Press, 1999).

[27] J. B. Campbell. *Introduction to Remote Sensing*, 2nd edition (The Guilford Press, New York, 1996).

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

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

[30] P. E. Gill, W. Murray, and M. H. Wright. *Numerical Linear Algebra and Optimization*, volume 1, p. 223 (Addison-Wesley Publishing Company, 1991).

[31] P. Groves and P. Bajcsy. "Methodology for hyperspectral band and classification method selection." In *Proceedings of IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data*. Greenbelt, MD (October 2003).

[32] P. Bajcsy and P. Groves. "Methodology for hyperspectral band selection." *Journal of Photogrammetric Engineering and Remote Sensing* (2003).

[33] G. Vane. "Airborne visible/infrared imaging spectrometer (AVIRIS): Description of the sensor, ground data processing facility, laboratory calibration, and first results." Technical Report JPL Publication 87-38, Jet Propulsion Lab, NASA, Pasedena, California (November 1987).

[34] C. Wessman. "Hyperspectral imagery with gramma labels." (October 1999).

[35] D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*, 2nd edition (Champman & Hall/CRC, Boca Raton, FL, 2000).

[36] P. H. Swain and S. M. Davis. *Remote Sensing: The Quantitative Approach* (McGraw-Hill, New York, 1978).

[37] C. A. Balanis. *Advanced Engineering Electromagnetics* (John Wiley and Sons, USA, 1989).

[38] "Veris technologies." (2003).