070390

HIERARCHICAL CLASSIFIER DESIGN IN HIGH DIMENSIONAL, NUMEROUS CLASS CASES

B. Kim and D. A. Landgrebe[1]

School of Electrical Engineering
Purdue University
West Lafayette, IN 47907, U.S.A.
Tel: (317)494-3486, FAX: (317)494-6440

ABSRACT

As the progress of new sensor technology continues, increasingly high resolution imaging sensors are being developed. These sensors will provide more detailed, complex data for each picture element and greatly increase the dimensionality of data over past systems. As a result, they should make possible discrimination between a much larger number of classes.

In applying pattern recognition methods in remote sensing problems, an inherent limitation is that there is almost always only a small number of training samples with which to design the classifier. The growth in both the dimensionality and the number of classes is likely to aggravate this already significant limitation of training set size. Thus ways must be found for future data analysis which can perform effectively in the face of large numbers of classes without unduly aggravating the limitations on training.

In this work we propose a hybrid decision tree classifier design procedure which produces efficient and accurate classifiers for this situation. In doing so, several key questions are addressed. Remote sensing systems which will perform pattern recognition tasks on high dimensional data with small training sets require efficient methods for feature extraction and prediction of the optimal number of features to achieve minimum classification error. Three feature extraction techniques are used here. Canonical and extended canonical techniques are mainly dependent upon the mean difference between two classes. An autocorrelation technique is dependent upon the correlation differences. The mathematical relationship between sample size, dimensionality, and risk value is derived. The incremental error is simultaneously affected by two factors, dimensionality and separability. Empirical results indicate that a reasonable rule of thumb for sample size is six to ten times the dimensionality. Empirical tests comparing the hybrid design classifier with a conventional singly layered one are also presented. They suggest that the hybrid design produce higher accuracy with fewer features. The need for fewer features is an important advantage, because it reflects favorably on both the size of the training set needed and the amount of computation time that will be needed in analysis

---

[1] Corresponding Author

I. INTRODUCTION

As progress in new sensor technology for Earth observational remote sensing continues, increasingly high spectral resolution multispectral imaging sensors are being developed. HIRIS (High Resolution Imaging Spectrometer) [1], for example, will gather data simultaneously in 192 spectral bands in the 0.4 - 2.5 micrometer wavelength region at 30 m spatial resolution. AVIRIS (Airborne Visible and Infrared Imaging Spectrometer) covers the 0.4 - 2.5 micrometer in 224 spectral bands. These sensors give more detailed, complex data for each picture element and greatly increase the dimensionality of data over past systems.

As a result, the new data should make possible discrimination between a much larger number of classes. If one envisions the hierarchy of information that might be derived in the form of an inverted information tree with broad general classes at the top, subdivided to ever more detailed classes as one proceeds downward, the new data should provide the possibility for a significantly deeper penetration into the information tree to more detailed and subtle classes. On the other hand, the typical practical situation with regard to data analysis of having only a limited amount of design data in the form of training samples by which to define the classes is not likely to change over past cases. It is thus clear that new types of analysis algorithms which will function effectively at much higher dimensionality with many more classes but no more design data will be needed if the full potential of the new sensors is to be achieved.

A useful approach in the face of such increased information complexity is to utilize a hierarchy or taxonomy in the analysis process, just as such a hierarchy is used in cataloging or displaying information. In the analysis process the potential advantages are increases in accuracy, speed, and level of detail which can be reached in the analysis process in that the analysis process at any given node of the tree can be more directly focused on the particular decision process needed at that node and on a more relevant subset of the data. Decision Tree Classifiers (DTC's) have been under study for some time in various application areas. Examples are remote sensing, character recognition, and blood cells classification. We next survey briefly some relevant aspects of these past works.

Wu et al.[2] defined a tree design process using an evaluation function to obtain the optimal DTC. Wu et al [2] and Swain and Hauska [3] suggested a histogram approach to decision tree structure design. However, since this approach used only one feature, the inter-relationships with other features are disregarded. Kanal [4] defined two types of admissible search strategies to obtain the optimal decision tree structure, namely S-admissible and B-admissible which have the cost of path and risk function in a state space graph model. Using the above functions, it is impossible to evaluate successfully every combination of tree structure to determine the overall optimal tree classifier in high dimensional, numerous class cases. You and Fu [5] designed binary trees by splitting a set of classes into two non-overlapping subgroups at every node. The two subgroups are found by comparing a measure of separability for different pairs of subgroups over various subsets of feature space with a fixed number of features. Landeweered et al.[6] suspected that binary tree classifiers improved the correct recognition rate compared with the application of single layer classifiers.

Casey and Nagy [7] developed a binary tree for optical character recognition using an information theoretic approach. The effectiveness of a node-by-node design scheme is highly dependent on the rule by which pixels are evaluated for assignment to a given node. The first

pixel to be tested is predetermined for the root node. A measure based on entropy is used for a pixel selection criterion. The rule employed for pixel selection is to choose the pixel that minimizes entropy, i.e, the one that maximizes the information gain. A priori class probabilities and class conditional frequencies of individual pixels are estimated from labeled samples. Their approach is a special case of binary tree character recognition.

In the ideal case feature selection should be simultaneously considered with decision tree structure considerations. Practically, Swain and Hauska [3] chose a feature selection criterion based on pair-wise separability over all pairs of classes after designing the decision tree structure. Muasher and Landgrebe [8] experimentally studied an effective feature ordering technique in cases which the number of training samples was limited in classifying two-class multivariate normal distributions.

Classifiers are usually designed with a finite sets of samples, and the estimation of the class-conditional densities which determine the decision boundaries is based upon these samples. In this case, the performance of the classifier is observed to improve up to a certain point as additional features are added, then deteriorate [9]. This is referred to as the Hughes phenomenon. D. H. Foley [10] investigated the design set error rate for a two class problem with multivariate normal distributions, and derived it as a function of the sample size per class and dimensionality. The design set error rate is compared to both the corresponding Bayes error rate and test set error rate. It was shown that the design-set error rate is biased below the true error rate and the test-set error rate is biased above the true error rate of classifier when the ratio of sample size to feature size is small.

Because of the Hughes phenomenon one needs to know how many features one should use at each node to maximize the overall classification accuracy. The number of features, the number of samples, and the correct classification accuracy are interrelated in a complex fashion. In the case of limited training samples and multidimensional space, the estimates of the first and second order statistics cannot accurately depict all the information which is contained in the data. In particular, the estimate of the covariance matrix may be poor. Therefore how to relate the inaccuracy of the estimate with classification error directly is important. Statistical distance measures are commonly used for this purpose. A. K. Jain [11] showed that when features have multinomial or univariate Gaussian distributions, the estimate of Bhattacharyya distance is biased and consistent. The bias and the variance of the estimate is not only a function of the number of training samples but also depends on the true parameters of the densities. Raudys [12] and Fukunaga [13] showed that the required number of training samples is proportional to the dimensionality to achieve a certain amount of error for a linear classifier, and is proportional to the square of dimensionality for quadratic classifier.

The design process of Decision tree classifiers (DTC) may be thought of as containing three elements: the tree structure design, the decision rule selection, and feature selection. To be truly optimal, these may not be pursued sequentially but must be simultaneously accomplished, and at all nodes at once. This design task in its full generality is thus complex, and simplifying restrictions and assumptions particularly tailored to the intended application are usually necessary to make the problem tractable. We will restrict our considerations here to binary trees (decisions between only two classes at each node) and assume a maximum likelihood classification of Gaussian classes and subclasses.

A key matter in the design process of any classifier is the definition of classes. A set of requirements for a valid list of classes for remote sensing data is that:

- The classes must each be of informational value (i.e. useful in a pragmatic sense).
- The classes must be spectrally or otherwise separable (i.e., distinguishable based on the available data).

Note that the former requirement derives from the intended application, while the latter rests on characteristics of the data. Thus a key consideration in the proper design of a classifier is how to reconcile a property of data, separability, with a property of the application, class informational value.

Summarizing then, the key elements of the problem are as follows.

- The increased dimensionality of future Earth observational space sensors should make possible analysis into a much larger number of more subtle and detailed classes,

- Decision Tree Classifiers are a useful way of dealing with the much larger number of classes, if an objective and straightforward means can be found for designing the decision tree,

- A characteristic of remote sensing analysis problems is that there is nearly always a sparsity of design data or training samples, since significant effort is usually required to determine and label these samples.

- Due to the increased dimensionality of the new data and the Hughes phenomenon, the sparse training sample problem will be aggravated, making feature determination including determination of the optimum dimensionality of increased importance.

- The tree design procedure must provide a suitable means for reconciling the requirement for the design procedure to result in separable classes with that of the final classes being of informational value.

Before proceeding with definition of a DTC design procedure, we will consider the issues of feature extraction and of estimation of the optimal dimensionality.

II. FEATURE EXTRACTION

A. Hughes Phenomenon

As previously stated, there is an optimum feature dimensionality to use relative to each set of classes and training set size, as predicted by Hughes [9]. If there were no such dimensionality phenomenon, the single layer maximum likelihood classifier would have better performance than the any other DTC because the conventional Bayes classifier gives the minimum classification error. However, when the number of training samples are limited,

the Hughes phenomenon must be considered. In such cases, the conditional density functions are inaccurately estimated because of the lack of adequate training samples. The poor estimates cause complex decision boundary to be biased, and obviously, this phenomenon is aggravated as the dimensionality of the data increases.

A common feature determination procedure is to choose a subspace by calculating the pair-wise Bhattacharyya distance at each node, then selecting the subsets of features having the largest distance for dimensionality reduction. Since the estimated means and covariances themselves are randomly biased in the limited sample situation, a better way to determine the features to be used is desirable. A feature design procedure which reduces the dimensionality while maintaining or even enhancing the separability would thus be very desirable.

B. Canonical Analysis

Fisher's suggestion [14] was to look for the linear function which maximizes the ratio of the between-class scatter to the within-class scatter. Canonical analysis finds a linear combination of the variables such that values are as close as possible within classes and as far apart as possible between classes. In canonical analysis, within-class and between-class scatter matrices are used to formulate a criterion of class separability. A within-class scatter matrix shows the scatter of samples around their class mean vector $\mathbf{m}_i$, and is expressed by

$$\mathbf{S}_w = \sum_{i=1}^{m} P(w_i)\, E\{(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T | w_i\} = \sum_{i=1}^{m} P(w_i)\, \Sigma_i \qquad (1)$$

The matrix $\mathbf{S}_w$ is proportional to the sample covariance matrix. A between-class scatter matrix is given by

$$\mathbf{S}_b = \sum_{i=1}^{m} P(w_i)\, (\mathbf{m}_i - \mathbf{m}_o)(\mathbf{m}_i - \mathbf{m}_o)^T \qquad (2)$$

$$\mathbf{m}_o = E\{\mathbf{x}\} = \sum_{i=1}^{m} P(w_i)\, \mathbf{m}_i \qquad (3)$$

where $\mathbf{m}_i$ is the mean of the i^{th} class and $\mathbf{m}_o$ is the global mean. All these scatter matrices are invariant under coordinate shifts. We define the ratio of the between class scatter to the within class scatter as:

$$\frac{\mathbf{d}^T \mathbf{S}_b \mathbf{d}}{\mathbf{d}^T \mathbf{S}_w \mathbf{d}} \qquad (4)$$

If $\mathbf{d}$ is the vector which maximizes the ratio, $\mathbf{d}^T\mathbf{x}$ is called the Fisher's linear discriminant function or the first canonical variate.

The eigenvector $\mathbf{d}_i$ which corresponds to eigenvalue λ_i is directly obtained from $\frac{\mathbf{S}_b}{\mathbf{S}_w}$. The eigenvector $\mathbf{d}_i$ is called the i^{th} canonical variate. If there are only two classes, the ratio has only one nonzero eigenvalue. The other *n-1* features do not contribute to the ratio. The final solution for two classes is

$$\mathbf{d} = \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \tag{5}$$

This is also called Fisher's linear discriminant function which has the maximum ratio of between-class scatter to within-class scatter.

C. Extended Canonical Analysis

The following method has been developed by Foley and Sammon [15]. In the two class problem, Fisher's vector is given by $\mathbf{d}_1 = \frac{\alpha_1(\mathbf{m}_1 - \mathbf{m}_2)}{\mathbf{S}_w}$ where α_1 is chosen such that $\mathbf{d}_1^T\mathbf{d}_1 = 1$. The next best direction can be found for maximizing the Fisher criterion subject to the constraint that $\mathbf{d}_1$ and $\mathbf{d}_2$ are orthogonal. Using the the method of Lagrange multipliers, we wish to maximize the Fisher criterion subject to the constraints that $\mathbf{d}_i^T\mathbf{d}_n = 0$ for i = 1,2,...,n-1. A recursive definition for the n^{th} discriminant vector is

$$\mathbf{d}_n = \alpha_n \mathbf{S}_w^{-1}\left\{(\mathbf{m}_1 - \mathbf{m}_2) - [\mathbf{d}_1 \cdots \mathbf{d}_{n-1}]\mathbf{S}_{n-1}^{-1}\begin{bmatrix}\frac{1}{a_1}\\ 0\\ \vdots\\ 0\end{bmatrix}\right\} \tag{6}$$

D. Autocorrelation Analysis

The Fisher concept can be applied to an autocorrelation matrix. In a two class problem, the criterion which maximizes $\mathbf{S}_1$ and minimizes $\mathbf{S}_2$ simultaneously or vice versa can be defined. An autocorrelation matrix is defined by

$$\mathbf{S}_i = \Sigma_i + \mathbf{m}_i\mathbf{m}_i^T \tag{7}$$

The criterion function is defined by $r = \frac{\mathbf{d}_i^T\mathbf{S}_1\mathbf{d}_i}{\mathbf{d}_i^T\mathbf{S}_2\mathbf{d}_i}$ or $\frac{\mathbf{d}_i^T\mathbf{S}_2\mathbf{d}_i}{\mathbf{d}_i^T\mathbf{S}_1\mathbf{d}_i}$. The optimally separable feature set is a feature set such that $\mathbf{S}_1$ is minimized and $\mathbf{S}_2$ is maximized or vice versa after the transformation. The ratio *r* is maximized by the selection of feature d if $\frac{\partial r}{\partial \mathbf{d}} = 0$. That equation can be reduced to $(\mathbf{S}_1 - r\mathbf{S}_2)\mathbf{d} = 0$ which is called a generalized eigenvalue equation.

$$(\mathbf{S}_2^{-1}\,\mathbf{S}_1 - \mathbf{R})\mathbf{d} = 0 \tag{8}$$

$$\mathbf{S_2}^{-1}\,\mathbf{S_1}[\mathbf{d}_1 \ldots \mathbf{d}_q] = \mathbf{R}[\mathbf{d}_1 \ldots \mathbf{d}_q] \tag{9}$$

We can diagonalize two symmetric matrices as

$$\mathbf{D}^T\mathbf{S_1}\mathbf{D} = \mathbf{I} \tag{10}$$

$$\mathbf{D}^T\mathbf{S_2}\mathbf{D} = \mathbf{R} \tag{11}$$

where **D** and **R** are the eigenvector and eigenvalue matrices of $\mathbf{S_1}^{-1}\mathbf{S_2}$. To find the orthonormal eigenvectors of $\mathbf{S_1}^{-1}\mathbf{S_2}$ as

$$\mathbf{S_1}^{-1}\,\mathbf{S_2}\mathbf{d}_i = r_i\mathbf{d}_i \quad \text{and} \quad \mathbf{d}_i^T\mathbf{d}_j = \delta_{ij} \tag{12}$$

We change the scale of $\mathbf{d}_i$ to satisfy

$$\alpha^2\mathbf{d}_i^T\mathbf{S_1}\mathbf{d}_i = 1 \tag{13}$$

Therefore, the i^{th} orthonormal eigenvector is

$$\mathbf{d}_i' = \frac{\mathbf{d}_i}{(\mathbf{d}_i^T\mathbf{S_1}\mathbf{d}_i)^{1/2}} \tag{14}$$

For each discriminant vector $\mathbf{d}_i'$, there corresponds an r_i,, given by

$$r_i = \frac{\mathbf{d}_i'^T\mathbf{S_1}\mathbf{d}_i'}{\mathbf{d}_i'^T\mathbf{S_2}\mathbf{d}_i'} \tag{15}$$

Each r_i represents the value of the discriminatory criterion for the corresponding discriminant vector $\mathbf{d}_i'$. The discriminant vectors can be ordered according to their respective ratio values such that

$$r_1 \geq r_2 \geq \ldots \geq r_q \geq 0 \tag{16}$$

However, we want to maximize the relative ratio between $\mathbf{S_1}$ and $\mathbf{S_2}$ which is greater than one. If an r_i is less than one, we should use the inverse value of r_i to compare the relative ratio. We may define the relative ratio as follows:

$$r_1 \geq r_2 \geq \ldots \geq r_i \geq 1$$

$$\frac{1}{r_q} \geq \frac{1}{r_{q-1}} \geq \ldots \geq \frac{1}{r_{i+1}} \geq 1 \tag{17}$$

The best feature or most effective basis function for both classes is the eigenvector corresponding to the largest relative ratio. The autocorrelation analysis can be used in place of canonical analysis when the mean difference between two classes is almost zero.

When the mean difference is zero the canonical analysis and the extended canonical analysis can not be used since the feature vector is defined by the mean difference. The autocorrelation analysis is useful when the mean difference is small and the covariance difference is dominant while the canonical analysis and the extended canonical analysis are more effective than the autocorrelation analysis when the mean difference is dominant. After extracting the feature, the mean difference and the covariance difference in the subspace are checked to choose between the two methods, extended canonical analysis or autocorrelation analysis, for the next feature extraction.

III. ESTIMATION OF OPTIMAL NUMBER OF FEATURES

As previously stated, it is well known that classifier accuracy expressed as a function of the number of features used shows a maximum at some finite dimensionality [9], and for a given class-conditional density function set, the occurrence of this peak is dependent upon the training sample size [16], as a result of the accuracy dependence upon the quality with which the density parameters are estimated.

A long-known fundamental barrier to the optimal design of classifiers is the inability to be able to directly calculate the expected accuracy of a trial classifier design. As a result, a common practice is to use a statistical measure, e.g. Bhattacharyya distance, to estimate the expected accuracy. However the relationship of such distance measures to classification accuracy, though monotonic, is not precisely one-to-one. Thus, if such a distance is to be used in the design process, it is important to clearly understand just what the relationship is between expected accuracy and a specific distance measure used to estimate it. It is this relationship which is studied next, paying specific attention to the effects of sample size and parameter estimation variability.

A. Optimal Number of Features

The risk function of an estimate T(x) is defined by

$$R(\theta, T) = E_{\theta}^{X}[L(\theta, T(x))] = \int_{-\infty}^{\infty} L(\theta, T(x))\, dF(x|\theta) \tag{18}$$

where $L(\theta, T(x))$ is the loss function, θ is a real parameter, and T(x) is an estimate. One may choose the mean squared error as the loss function such that $L(\theta, T) = (\theta - T)^2$. Then

$$R(\theta, T) = E\left[\left(T(x) - q(\theta)\right)^2\right] = Var(T(x)) + \left[E\left(T(x) - q(\theta)\right)\right]^2 \tag{19}$$

Usually, it is difficult to obtain the risk value of a functional directly. Therefore, Taylor series expansion techniques may be applied to approximate the risk value of the functional. The Bayes error ε^* can be expressed by

$$\varepsilon^* = \int_{-\infty}^{\infty} \min[P_1 p_1(x)\ P_2 p_2(x)]\, dx \tag{20}$$

If the class-conditional density functions are Gaussian, then

$$\varepsilon^* \leq \sqrt{p_1 p_2}\, \exp[-B] \tag{21}$$

where

$$B = \frac{1}{8}(m_1 - m_2)^T \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1} (m_1 - m_2) + \frac{1}{2} \ln \frac{\left|\frac{\Sigma_1 + \Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}} \tag{22}$$

Here B is the Bhattacharyya distance for the Gaussian case, $\mathbf{m}_i$ is the mean vector of class i, and Σ_i is the covariance matrix of class i. Note that the first term of the Bhattacharyya distance reflects the separation due to the mean difference between the two classes, and the second term reflects the covariance difference. The Bhattacharyya distance measure usually bears a closer relationship with the classification accuracy than other measure functions such as divergence [18].

If one assumes that Bayes error is directly related to Bhattacharyya distance, the estimated Bhattacharyya distance behavior can show that the increment of Bayes error comes from the poorly estimated Bhattacharyya distance. However, the Bayes error is not bounded by the Bhattacharyya distance but merely a function of Bhattacharyya distance.

We may define the transformed Bhattacharyya distance as:

$$X_B = 1 - \sqrt{p_1 p_2}\, \exp[-B] \tag{23}$$

The transformed Bhattacharyya distance is assumed to be directly related to the classification accuracy. If one assume that the Bayes error is approximately equal to the upper bound that is characterized by Bhattacharyya distance, then

$$\varepsilon^* \approx \sqrt{p_1 p_2}\, \exp[-B] \tag{24}$$

The transformed Bhattacharyya distance is the lower bound of the correct classification accuracy. If P_1 and P_2 are equal, the estimated error of the classifier designed by training samples can be expressed as

$$\hat{\varepsilon} \approx \frac{1}{2}\, e^{-\hat{B}} \tag{25}$$

In the multivariate statistical analysis, a powerful property is that the Bhattacharyya distance is invariant under any one-to-one mapping. By simultaneous diagonalization,

$$\mathbf{m}^{(1)} = \mathbf{0},\ \mathbf{m}^{(2)} = \mathbf{m},\ \Sigma^{(1)} = \mathbf{I},\ \Sigma^{(2)} = \mathbf{A} \tag{26}$$

The number of parameters for the estimated Bhattacharyya distance is $2(q + q^2)$ where q is the dimensionality. The bias of the estimated Bhattacharyya distance is derived in [11][13]. For the computation of the derivatives of the Bhattacharyya distance-containing matrix, three basic matrix differential equations are needed.

$$\frac{\partial \Sigma^{-1}}{\partial \lambda_{ij}} = -\Sigma^{-1} U_{ij} \Sigma^{-1} \tag{27}$$

$$\frac{\partial |\Sigma|}{\partial \Sigma} = |\Sigma| (\Sigma^{-1})^T \tag{28}$$

$$\frac{\partial \Sigma^T A \Sigma}{\partial \lambda_{ij}} = U_{ij} \Sigma^T + \Sigma U_{ij}^T \tag{29}$$

where U_{ij} matrix has all zero valued components except that the i^{th} column and j^{th} row component is one.

The bias and the variance of the transformed Bhattacharyya distance can be obtained as follows where q is the dimensionality, n is the number of samples, and B is the Bhattacharyya distance.

We want to find the risk value of the transformed Bhattacharyya distance because increasing the risk value makes the classification error increase. The risk function of the transformed Bhattacharyya distance is

$$R(\hat{X}_B, X_B) = \frac{1}{4}\left(E\left[e^{-B} - e^{-\hat{B}}\right]\right)^2 + \mathrm{Var}\left[1 - \frac{1}{2}e^{-\hat{B}}\right] \approx \frac{1}{4}\left(E\left[e^{-B} - e^{-\hat{B}}\right]\right)^2 \tag{30}$$

$$E\left[\frac{1}{2}e^{-B} - \frac{1}{2}e^{-\hat{B}}\right] \approx \frac{e^{-B}}{8n}\left[q - \sum_{i=1}^{q} \frac{m_i^2}{(1+\lambda_{ii})} + \sum_{i=1}^{q}\sum_{j=1}^{q} \left(\frac{m_j^2(1+\lambda_{ii}\lambda_{ij})}{(1+\lambda_{jj})^2(1+\lambda_{ii})} - \frac{(1+\lambda_{ii}\lambda_{ij})}{4}\left(\frac{m_i\, m_j}{(1+\lambda_{ii})(1+\lambda_{jj})} \right)^2 \right) \right.$$
$$\left. + \sum_{i=1}^{q} \left(\frac{m_i^2(1+\lambda_{ii}^2)}{(1+\lambda_{ii})^3} - \left(\frac{1}{(1+\lambda_{ii})} - \frac{1}{2} - \frac{m_i^2}{2(1+\lambda_{ii})^2} \right)^2 \right) + \sum_{i=1}^{q} \lambda_{ii}^2 \left(\frac{1}{(1+\lambda_{ii})} - \frac{1}{2\lambda_{ii}} - \frac{m_i^2}{2(1+\lambda_{ii})^2} \right)^2 \right]$$
$$+ \frac{e^{-B}}{8n}\left[q(q+1) - \sum_{i=1}^{q}\sum_{j=1}^{q} \frac{(1+\lambda_{ii}\lambda_{ij})}{(1+\lambda_{ii})(1+\lambda_{jj})} - \sum_{i=1}^{q} \frac{(1+\lambda_{ii}^2)}{(1+\lambda_{ii})^2} \right] \tag{31}$$

$$\mathrm{Var}\left[1 - \frac{1}{2}e^{-\hat{B}}\right] \approx \frac{e^{-2B}}{16n}\left[\sum_{i=1}^{q} \frac{m_i^2}{(1+\lambda_{ii})} + \sum_{i=1}^{q}\sum_{\substack{j=1 \\ i \neq j}}^{q} \frac{m_i^2 m_j^2 (1+\lambda_{ii}\lambda_{jj})}{2(1+\lambda_{ii})^2(1+\lambda_{jj})^2} \right.$$

$$+\sum_{i=1}^{q} 2\left(\frac{1}{(1+\lambda_{ii})} - \frac{1}{2} - \frac{m_i^2}{2(1+\lambda_{ii})^2}\right)^2 + \sum_{i=1}^{q} 2\lambda_{ii}^2\left(\frac{1}{(1+\lambda_{ii})} - \frac{1}{2\lambda_{ii}} - \frac{m_i^2}{2(1+\lambda_{ii})^2}\right)^2\Bigg] \qquad (32)$$

From (31) note that the exponential term, $\frac{e^{-B}}{8n}$, implies a reduction of risk as q increases since B increases with q, while the rest of the terms cause the value of the risk to increase.To minimize the risk value with a constraint to maximize the Bhattacharyya distance, the dimensionality must be reduced when the number of training samples is small.

If the features are ordered, the first feature reduces the risk function and expands the Bhattacharyya distance the most. As the number of features is increased, the summation terms increase more rapidly than the exponential term. The strategy for the prediction problem can be established as follows. If one wants to use as small a number of features as possible and achieve as a large Bhattacharyya distance as possible, one should take advantage of the transformed coordinates. The best one feature having the smallest risk value and largest Bhattacharyya distance between any two classes can be extracted in the transformed coordinates in the case of a small training sample situation.

It may be noted from equation (31), that if only the mean difference term is considered as in the case of a linear classifier with a fixed Bhattacharyya distance, the bias increases linearly with the dimensionality, while if the both mean and covariance terms are considered with a fixed separability for the case of a quadratic classifier, the bias increases quadratically with the dimensionality. These results agree with previous works [12,13].

Since the separability usually increases as dimensionality increases, the incremental error is dependent upon not only dimensionality but separability. In the next section, the required sample size is studied empirically for general cases.

B. Empirical Study

Fukunaga and Krile [16] developed an algorithm for calculating recognition error when applying pattern vectors in an optimum Bayes' classifier. When the q random variables of the vector **x** are independent, the characteristic function of h(**x**) for class *i* is

$$\phi_i(w) = E\{e^{jwh(\mathbf{x})} \mid \text{class } i\} = \int_{-\infty}^{\infty} e^{jwh(\mathbf{x})} p_i(\mathbf{x})\, d\mathbf{x} = \sum_{l=1}^{q} \int_{-\infty}^{\infty} e^{jwh(x_l)} p_i(x_l)\, dx_l \qquad (33)$$

Of course, once the characteristic function of h(**x**) is obtained the density function of h(**x**) is found by use of the inverse Fourier transform.

$$p(h|\text{class } i) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_i(w)\, e^{-jwh}\, dw \qquad (34)$$

Since two covariance matrices can be diagonalized simultaneously by linear transform, when the distributions are normal, all features in the transformed coordinate system are independent. The errors are invariant under any transformation because the likelihood ratio is independent of any coordinate system. Characteristic functions of the minus log likelihood ratio for class 1 and class 2 can be easily computed because the q random variables of vector **x** are independent. This approach reduce the q dimensional integral to a one dimensional integral for the error from each class.

$$\varepsilon^* = P_1 \int_0^{\infty} p_1(h)\, dh + P_2 \int_{-\infty}^{0} p_2(h)\, dh \tag{35}$$

To examine the global relationships between the dimensionality, the sample size, and the correct classification accuracy, Monte Carlo simulation is used here. True Bayes' error can be computed numerically by Fukunaga's algorithm if one has perfect knowledge of the mean and covariance with Gaussianly distributed classes. Although only 1-dimensional numerical integration is needed for Fukunaga's algorithm, it is difficult to obtain an accurate Bayes' error estimate easily in high dimensionality. Therefore, a simpler means to estimate the Bayes' error is needed to study relationships between sample size, dimensionality, and added error empirically. Whitsitt and Landgrebe [18] suggested that if we let f = erf, then we are assured that the locus of (p_e, f) contains p_e = f, and in this sense, f approximates error.

The error function Bhattacharyya distance is defined by

$$E = 0.5 - 0.5\mathrm{erf}(\sqrt{B}) \tag{36}$$

The error function transformed Bhattacharyya distance is defined by

$$E_B = 1 - E = 0.5 + 0.5\mathrm{erf}(\sqrt{B}) \tag{37}$$

where the error function is given by

$$\mathrm{erf}(\sqrt{B}) = \int_{\sqrt{B}}^{\infty} \frac{\exp[-\frac{x^2}{2}]}{\sqrt{2\pi}}\, dx$$

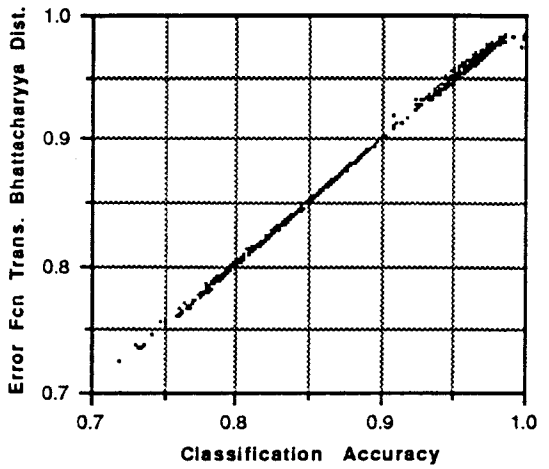

Fig. 1. Simulation Result for P_c vs E_B ($q = 10$, $n = \infty$)

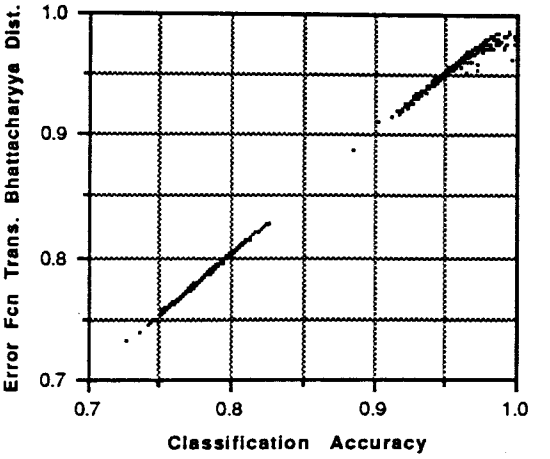

Fig. 2. Simulation Result for P_c vs E_B ($q = 30$, $n = \infty$)

Classification accuracy versus the error function transformed Bhattacharyya distance (E_B) is plotted for a dimensionality of ten in Fig. 1 and thirty in Fig. 2. Here the classification accuracy is obtained by Fukunaga's algorithm. Note that the probability of correct classification and the error function transformed Bhattacharyya distance have a linear relationship.

Therefore, E_B is selected to study the relationships between sample size, dimensionality, classification results and to observe the Hughes phenomenon to determine the optimal number of features in two class cases.

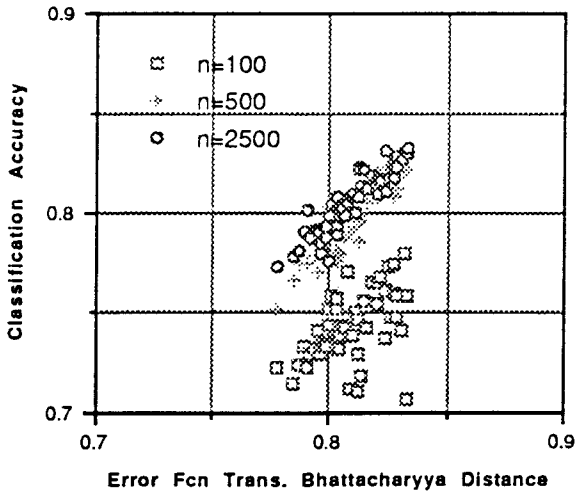

Fig. 3. Classification Accuracy vs E_B ($q = 50$)

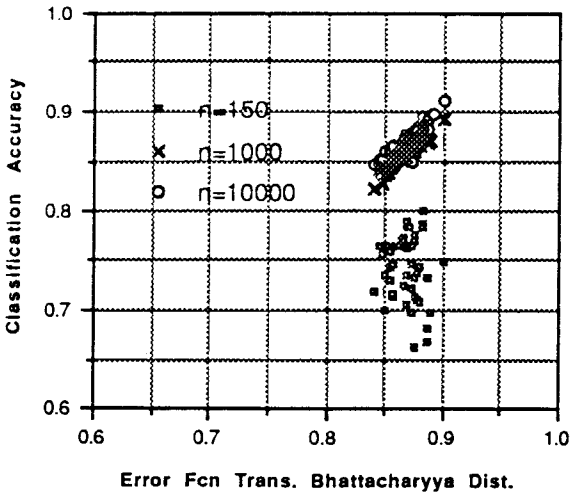

Fig. 4. Classification Accuracy vs E_B ($q = 100$)

The Hughes phenomenon is clearly shown in the simulation results shown in Fig. 3 and 4. The classification accuracy begins to fall below the 45° diagonals of Figures 1 and 2 as n, the

number of training samples used, is decreased from some large number. The number of simulations used here is fifty for a given number of training samples.

In these tests, about two times and ten times the dimensionality, and the power of two of the data dimensionality were used to estimate the class-conditional densities in the simulations. When two times the data dimensionality was used, the estimated classification accuracy was well below the true classification accuracy. When ten times the data dimensionality was used, the estimated classification accuracy is almost the same as the true classification accuracy. When the power of two of the data dimensionality was used, the estimated classification accuracy is similar to the result in the case of ten times the dimensionality of the data.

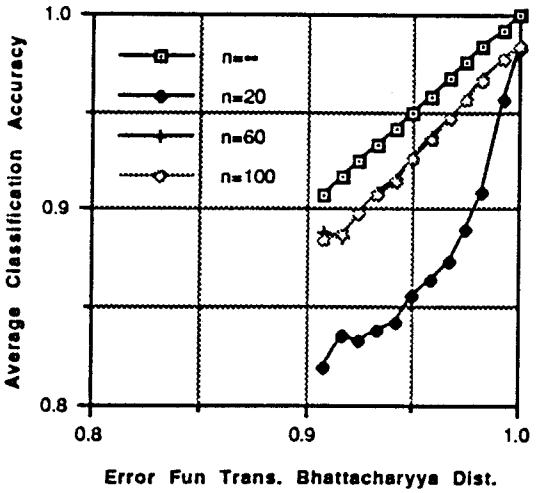

Fig. 5. Mean Value of P_c vs E_B (q = 10)

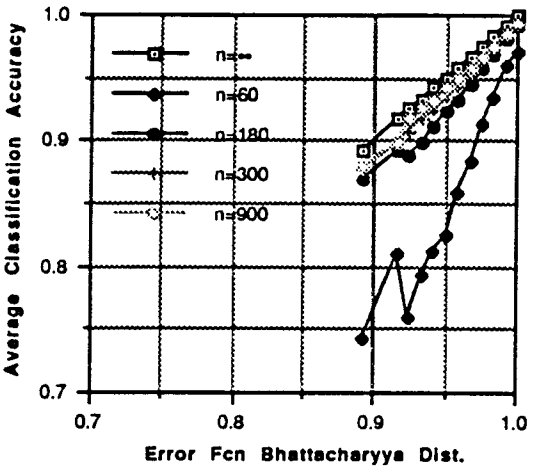

Fig. 6. Mean Value of P_c vs E_B (q = 30)

To illustrate the one-to-one relationship between estimated classification accuracy and E_B, the average classification accuracy is plotted in Fig. 5 and 6. From this, it appears that approximately six to ten times the number of training samples with respect to the dimensionality are needed to achieve a satisfactory design at this dimensionality.

As the dimensionality is increased, the separability is also increased since the added features give more information. When the dimensionality is increased with a fixed separability, the added error increases quadratically [2]. However, when the dimensionality and the separability increase together, it is difficult to find a simple relationship because the increased separability reduces the added error and the increased dimensionality cause the added error increase. Empirically, the cases of increasing both dimensionality and separability are tested.

IV. HYBRID DTC DESIGN

The previously suggested methods for DTC design, such as an evaluation function or admissible search, [2,4] are not feasible for the current problem since the complete conditional density functions are not available and a large amount of computation time would be needed to evaluate all combinations of tree classifier parameters. Practically, the methods

of minimizing the classification error at each node are implemented to obtain locally optimum results although the overall performance is not optimal [5].

As previously stated, to be a valid class, a distribution must be simultaneously of informational value and separable from the other classes. Supervised procedures can guarantee the former, but not the latter. Unsupervised procedures, e.g. clustering, can provide the latter, but do not guarantee the former. Thus, a practical classification scheme for DTC design must contain both procedures in such a way that the simultaneity of satisfaction is guaranteed.

More directly, there exist only alternatives which are a top down and a bottom up approach. The terminal classes must be both separable and of informational value. Non-terminal nodes are not required to be classes of informational value, but they must still be separable. Thus clustering, which insures separability, may be used in a top down approach, while correspondence with training sets is required at the bottom and thus is related to bottom up approach.

Although tree structure, decision rules and optimal feature sets should be simultaneously considered to obtain a globally optimal DTC, some compromise with tractability is required here. Given the parametric approach, the decision rule will be selected first, because the criterion for designing a DTC is determined by the decision rule. The binary tree is chosen as the type of DTC structure, because any tree can be reduced to a binary tree and the most effective feature subsets can be obtained for a binary tree. For the other requirements of data properties which are terminal classes of informational value and spectral separability, a hybrid design approach, which alternately proceeds bottom up and then top down, is proposed.

A. DTC Design

First, consider clustering, and in particular, the means for its initialization. The determination of the initial cluster centers is an important matter because clustering results differ depending on the initial cluster centers. Since there is usually more than one set of final clusters which exhibit adequate separability, one that is as close as possible to the set for the classes of information value is the most desirable. Further, the nearer the initial cluster centers to the final ones, the fewer the number of clustering iterations required.

It is thus logical to seek effective initial cluster centers by beginning from the bottom up, i.e., with the training data. First the separability between every class pair is computed. Then the two classes that has the smallest separation value are merged. This choice has the effect of maximizing the distance measure at the next upper level. Thus a new class, or subgroup, at that next level up the tree is created. Then the new Bhattacharyya distances are computed between the new subgroup and the remaining classes or subgroups. The above procedures are continued until all classes become two subgroups. The final two subgroups give two initial cluster centers and covariance information.

The subgroup information is used for the top down approach. The normalized sum of squared error(NSSE) criterion, defined as follows, is used.

$$\sum_{i=1}^{c} \sum_{x \in c_1} \left[(x - m_i)^T \Sigma_i^{-1} (x - m_i) + \ln|\Sigma_i| \right] \quad (38)$$

Note that with this normalized sum of squared error criterion, variance effects are considered.

The hybrid design proceeds as follows:

1. Divide the entire data set into two subgroups for the descendent nodes by the bottom up approach.

2. Compute the mean and covariance vectors of the two subgroups and re-divide the classes into two subgroups by the top down approach using the Normalized Sum of Squared Error Clustering.

3. If the separated groups are informational classes, go to step 4. Otherwise, return to step 1 for each subgroup which is not an informational class.

4. Design is complete

There are several advantages to the hybrid approach. It is more likely to converge to classes of informational value because the initialization provides early guidance in that direction while the straightforward top down approach does not guarantee such convergence. It can use overlapping classes while there are no overlapping classes in the bottom up approach. Covariance information can be applied in the hybrid approach to separate non-spherical subgroups.

B. DTC for Multisource Data

Modern data sets often include not only spectral data but may also include other types of data, such as forest type maps, ground cover class maps, radar data, and topographic information such as elevation, slope, and aspect data. These are called multisource data. Because the multisource data are often not suitably modeled by multivariate distributions, conventional multivariate classification methods often cannot be used satisfactorily in analyzing multisource data. The data are not necessarily in common units and therefore scaling problems may arise. Further, the data may not even be numerical. Several methods have been proposed to classify the multisource data.

Hutchinson [19] proposed ambiguity reduction techniques. If the data are classified based on one or more data sources, the remaining ambiguities from the results of the classification are resolved by other sources. The stacked vector approach which consists of all components of all data sources has also been used [20]. This method is straightforward and simple. However, the method is not applicable when the various sources cannot be modeled by the multivariate distributions. Swain, Richards and Lee [21] proposed a statistical based analysis. In general there may not be a simple relation between the user-desired information classes and the set of data classes available. It is one of the requirements of a multisource analytical procedure to devise a method by which inferences about information classes can be drawn from the collection of data classes. They defined a set of global membership functions that collect together the inferences concerning a single information class from all of the data

source. They used the global membership function in the nature of a discriminant function, so that a pixel is then classified according to the usual maximum selection rule. In that case, the inter-source independence assumption was made. However, that assumption is not fully satisfied in the case of real data.

In the DTC approach, each source may be considered separately, something not possible in a single layered scheme. The basic idea is that optimal source and classification rules are determined to minimize the classification error at each node. To separate the subgroups evaluation functions are defined as a function of minimum error and minimum overlap. The overlap is defined as,

$$d = \sum_{i=1}^{c} n_i - n \tag{39}$$

where n_i is the number of samples of class i. The evaluation function is given by

$$E_i = \sum_{j=1}^{c} P_e(j) + \alpha d \tag{40}$$

where α is weighting factor.

For a hybrid DTC with a Gaussian maximum likelihood rule, two initial subgroups can be obtained by the bottom up approach with respect to each source. The subgroup consists of more than one informational class. To obtain the new subgroups, the Normalized Sum of Squared Error is applied for two clusters with respect to each source. To determine the best source, the evaluation function for each source is computed by evaluating the results of two clusters. Every node has the appropriate source to minimize the evaluation function. If a subgroup is not a informational class, the hybrid design procedure is applied again to obtain two descendent nodes.

V. EXPERIMENTS

A performance comparison of a DTC and a single layer classifier, a DTC approach for multitype data, and the effects of the feature extraction method in DTC design will be presented. The Bayesian decision rule with an assumed 0-1 loss function and multivariate normal distributions is used as the decision rule in all experiments when classification is involved. The 0-1 loss function assigns no loss to a correct decision, and unit loss to any error. Thus, all errors are assumed equally costly.

Three kinds of data sets are used as follows: Flight Line C-1(hereafter referred to as FLC-1), Anderson River Field Spectrometer System(FSS). FLC-1 data were measured and recorded from an aircraft flight on June 28, 1966, at approximately 12:30 PM local time, at an altitude of 2600 feet above terrain in Tippecanoe County, Indiana. A spatially scanning radiometer with a 3 mrad. spatial resolution was used to obtain relative measurements of the energy reflected from the ground in twelve different wavelength bands from 0.40 to 1.00 μm.

The Anderson River data set consist of 11 bands of airborne multispectral scanner(A/B MSS) data , 4 bands of synthetic aperture radar imagery(Steep SAR and Shallow SAR) and digital terrain model information including digital elevation, slope and aspect (DEM, DSM and DAM respectively). The A/B MSS Anderson River data was obtained over a Canadian forest site (2.8 km by 2.8 km) on July 29, 1978 at an altitude 3100 meters above sea level. The spatial resolution was 7 meters. Steep Mode SAR data was measured on July 25, 1978 over the site at an altitude 6700 meters above sea level. The raw data resolution was 3 meters. Shallow Mode SAR data was obtained on July 31,1978 at an altitude 6400 meters above sea level.

Six sets of high spectral resolution field measurement data were taken over Williams County, North Dakota and Finney County, Kansas. These data were taken by the Field Spectrometer System (FSS) mounted in a helicopter. The spectral resolution was 0.02 μm for the interval from 0.4μm to 2.4μm.

Eight classes of FLC-1 data were selected as follows: Alfalfa, Corn, Oats, Red Clover, Soybeans, Wheat, Bare Soil, and Rye. Fifteen training samples for each class were chosen such that the training set size is only slightly larger than the number of spectral features, thus providing an extreme test. At least one more sample than the number of features is needed to avoid singular covariance matrices. A large number of samples is used to evaluate the classification accuracy.

In Fig. 7, the hierarchical classifier is compared to the single layer maximum likelihood Gaussian classifier. Note that the best performance was provided by the DTC, and at the lowest dimensionality, implying less computation.

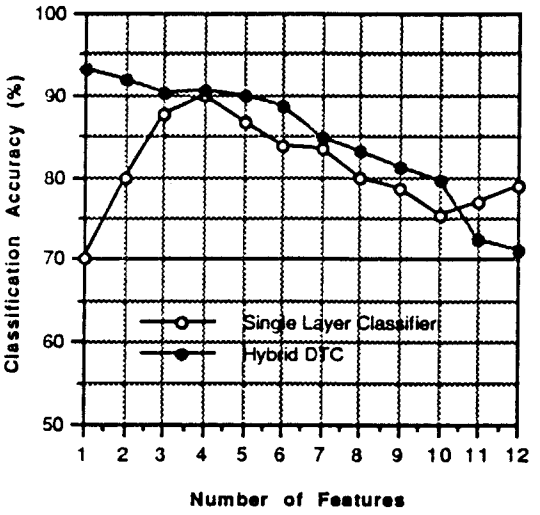

Fig. 7. Classification Accuracy vs Number of Features (8 Class)

Twenty three classes of FLC-1 data were chosen to test the performance of a hybrid DTC for the case of a larger number of less separable classes. The twenty three classes consist of two alfalfa fields, four corn fields, four oats fields, three red clover fields, five soybeans fields, three wheat fields, a bare soil, and a rye field. Fifteen training samples for each class were chosen and at least 492 samples were used for performance evaluate. To test the more complex data, the same species located on different areas, being a somewhat different state

of development, were considered as different classes. This procedure was chosen to provide a strong challenge to the classifier.

Fig. 8 shows that the the DTC again had better performance than the single layer classifier in this situation of a larger number of classes and a small number of features.

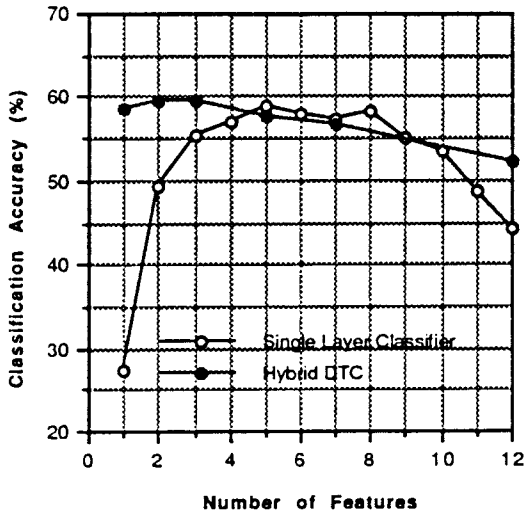

Fig. 8.Classification Accuracy vs Number of Features (23 Class)

To test the multisource, multitype data capabilities, the Anderson River data was used. Six classes were selected. Two subgroups for each source, A/B MSS, Steep SAR, Shallow SAR, DEM, DSM, and DAM, were obtained. Two subgroups provide the initial information which determines two initial cluster centers. After applying the normalized clustering algorithm for the top down approach, the best source is selected by comparing the evaluation function which is defined in equation (40). Thirty training samples for each class were used to estimate the parameters. More than 1200 samples were used to estimate the classification accuracy. The DTC reduced the error rate by 7.5 % over the single layer classifier with A/B MSS,as tabulated in Table 1.

Table 1. Multisource Data Result

Data Source	A/BMSS*	Steep SAR*	Shallow SAR*
Douglas-Fir	59.6	75.9	44.3
D-F + Lodgepole Pine	37.9	26.9	11.6
D-F + Cedar	43.1	18.9	20.1
Hemlock + Cedar	67.9	51.6	34.5
D-F + Other Species	0.1	74.6	15.3
Forest Clearings	5.1	67.7	24.1
Average(%)	57.0	52.6	25.0

Data Source	DAM*	DEM*	DSM*	DTC
Douglas-Fir	12.3	44.4	0	54.9
D-F + Lodgepole Pine	12.0	43.7	25.8	45.7
D-F + Cedar	82.1	0	0	43.4
Hemlock + Cedar	0	85.4	91.7	93.6
D-F + Other Species	29.0	95.9	43.2	86.9
Forest Clearings	0.5	22.2	17.4	62.3
Average(%)	22.6	48.6	29.7	64.5

Previously extended canonical analysis and autocorrelation analysis were introduced as feature extraction methods and the risk function of the classification accuracy was derived. To minimize the risk function with a constraint to maximize the Bhattacharyya distance, the dimensionality must be reduced while maximizing the Bhattacharyya distance when the number of training samples are small. Use of equation (36) showed that the best one feature in the transformed coordinate can give the best result in a situation which has only a small number of training samples. We note that the above statements are analytically definitive only for two classes. In the following experiments, we will test whether the best one feature which has the maximum separability does indeed produce the best performance.

FLC-1 data was used for the following experiment. In transformed coordinates, the first one feature was extracted by canonical analysis since the mean difference between classes was dominant. Next, features in a transformed subspace were extracted by extended canonical analysis or autocorrelation analysis because the mean difference became smaller in a subspace. Therefore, added features are obtained from the extended canonical analysis or autocorrelation analysis. In this experiment, the best one feature produces the best result here as shown in Table 2 and 3.

Table 2. Extended Canonical Result (FLC-1)

Num. of Feat.	1	2	3	4	5	6	7	12
Alfalfa	88.6	76.2	73.0	75.4	68.4	63.4	54.2	9.0
Corn	97.0	96.5	96.3	96.8	98.8	98.2	98.5	81.8
Oats	92.0	96.0	96.2	95.4	93.2	89.6	88.0	81.7
Clover	88.5	91.6	92.8	90.0	89.9	89.8	74.2	37.1
Bean	85.9	83.1	72.1	74.7	73.8	74.7	71.6	70.9
Wheat	99.4	99.2	99.2	99.4	99.2	99.0	98.2	98.8
Soil	99.7	99.7	99.9	99.9	99.9	100	100	94.8
Rye	95.5	93.5	93.7	93.4	93.3	94.0	94.4	93.2
Avg(%)	93.3	92.0	90.4	90.6	90.0	88.6	84.9	71.0

* The single layer classifier is applied.

Table 3. Canonical-Autocorrelation Result (FLC-1)

Num. of Feat.	1	2	3	4	5	6	7	12
Alfalfa	88.6	80.0	60.7	38.3	35.3	25.5	28.3	9.0
Corn	97.0	95.6	96.2	95.5	97.0	97.8	96.5	81.8
Oats	92.0	85.9	89.1	77.8	74.1	71.7	70.6	81.7
Clover	88.5	86.6	88.5	86.8	85.7	83.4	82.4	37.1
Bean	85.9	73.6	61.6	70.1	73.2	65.4	62.2	70.9
Wheat	99.4	99.8	99.4	99.6	99.6	99.6	99.6	98.8
Soil	99.7	100	100	99.9	99.9	99.9	99.9	94.8
Rye	95.5	96.9	99.4	98.6	98.7	96.9	98.4	93.2
Avg(%)	93.3	89.8	86.8	83.3	82.9	80.0	79.7	71.0

For a high dimensional data test, ten classes of FSS data were selected with forty training samples per class and more than 374 test samples for evaluating performance. The first one feature was extracted by canonical analysis since the mean difference between classes was dominant. Next, features in a subspace were extracted by extended canonical analysis. The smallest error rate was obtained for the best features as shown in Table 4. In this small training situation, the best one feature which had the largest separability in the transformed coordinate space provided the highest accuracy.

Table 4. Extended Canonical Result (FSS)

Num. of Feat.	1	2	3	4	5	10	20	30
Fallow1	63.5	59.2	57.4	57.1	56.4	57.9	47.3	24.9
Fallow2	82.6	83.4	83.4	82.1	81.8	82.4	67.9	36.1
Fallow3	67.5	68.5	68.5	68.3	71.3	58.7	57.7	69.0
Unknown1	41.3	44.1	44.6	48.1	45.2	44.2	49.8	59.2
Unknown2	71.0	69.2	68.9	66.6	66.5	58.9	55.8	36.0
Wheat1	56.5	56.8	56.8	54.2	57.0	61.2	59.9	76.1
Wheat2	83.5	83.4	82.6	83.5	81.8	81.5	79.1	88.2
Wheat3	85.3	84.7	84.7	84.0	82.8	76.5	66.6	59.5
Wheat4	49.4	50.8	50.0	48.7	50.8	50.5	47.1	35.4
Wheat5	83.8	83.7	83.9	83.3	81.8	74.4	67.5	81.1
Avg(%)	68.4	68.4	68.1	67.6	67.5	64.6	59.9	56.6

As a result, the hybrid DTC has better classification accuracy than the maximum likelihood Gaussian classifier when the best single feature is used at each node in the limited training sample situation.

VI. CONCLUDING REMARKS

The challenge of analyzing high dimensional data into a large number of classes with limited training sets is certainly a daunting one. The complexity of the problem requires that a suitable compromise be struck between the intellectually desirable goal of a globally optimal procedure and the pragmatically important one that robustly provides (a) near-optimal results (b) within an acceptable amount of computational effort, and (c) can ultimately be packaged in such a way as to be attractive and useful to the Earth scientist who will use it.

So far as DTC design, itself, the type of application in mind suggested focus upon devising a sound tree structure design procedure with adequate additional attention to efficient use of spectral features. Thus the key results presented here have to do with (a) adapting suitable feature transformation procedures from the literature, (b) developing a deeper understanding of the Hughes phenomenon, how it applies in this application, including the means for determination of what the optimal dimensionality is in a given case, and (c) devising of a hybrid of top down and bottom up DTC structure design methods so as to simultaneously satisfy the requirements for maximally separable classes at all levels of the tree and that of achieving the desired classes of informational value at the terminal nodes.

Though the results to this point represent, we believe, substantial progress toward the goal of providing an effective analysis tool for the coming era of Earth observational remote sensing, we do not wish to suggest that the task is now complete. The particular approach described here does appear from the results to have promise, and there are additional details regarding this work to be found in [22]. However, many more aspects remain to be investigated and understood. For example, the performance characteristics have only been explored in a limited set of circumstances; a wider set of tests and uses will surely raise additional matters which require further research.

ACKNOWLEDGEMENTS

We would like to thank the Canadian Centre for Remote Sensing, Department of Energy, Mines, and Resources, of the Government of Canada for allowing the use of the Anderson River Data. This work was funded in part by NASA under grant NAGW - 925.

REFERENCES

[1] A.F.H. Goetz and M. Herring, "The High Resolution Imaging Spectrometer (HIRIS) for EOS," IEEE Transactions on Geoscience and Remote Sensing, Vol. 27, No. 2, pp 136-144, March 1989.

[2] C. Wu, D. Landgrebe, and P. H. Swain, "The decision tree approach to classification," Tech. Rep. TR-EE 75-17, Purdue Univ., 1975.

[3] P. H. Swain and H. Hauska, "The decision tree classifier : Design and potential," IEEE Trans. Geoscience Electron., vol. GE-15 pp. 142-147, 1977.

[4] L. N. Kanal, "Problem-solving models and search strategies for pattern recognition," IEEE Trans. Pattern Anal. and Machine Intell., vol. PAMI-1, pp. 193-201, 1979.

[5] K. C. You and K. S. Fu, "An approach to the design of a linear binary tree classifier," Proc. Symp. Machine Processing of Remotely Sensed Data, Purdue Univ., 1976, pp. 3a-1 to 3a-10.

[6] G. H. Landeweerd, T. Timmers, and E. S. Gelsema, "Binary tree versus single level tree classification of white blood cells,"Pattern Recognition, vol. 16, pp. 571-577, 1983.

[7] R. G. Casey and G. Nagy, "Decision tree design using a probabilistic model," IEEE Trans. Inform. Theory, vol. IT-30, pp. 93-99, 1984.

[8] M. J. Muasher and D. A. Landgrebe, "The K-L expansion as an effective feature ordering technique for limited training sample size," IEEE Trans. Geosci. Remote Sen., vol. GE-21, pp. 438-441, 1983.

[9] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," IEEE Trans. Inform. Theory, vol. IT-14, pp. 55-63, 1968.

[10] D. H. Foley, "Consideration of sample and feature size," IEEE Trans. Inform. Theory, vol. IT-18, pp. 618-626, 1972.

[11] A. K. Jain, "On an estimate of the Bhattacharyya distance," IEEE Trans. Systems, Man, and Cyber. vol. SMC-6, pp. 763-766, 1976.

[12] S. Raudys, "On dimensionality, learning sample size and complexity of classification algorithms," Proc. Third Int. Joint Conf. Pattern Recognition, Coronado, CA, pp. 166-169, 1976.

[13] K. Fukunaga and R. A. Hayes, "Effect of Sample Size in Classifier Design ," IEEE Trans. Pattern Anal. Machine Intell., vol PAMI-11, pp. 873-885, 1989.
[14] R.O. Duda and P. E. Hart, *Pattern classification and scene analysis.* John Wiley, 1973.
[15] D. H. Foley and J. W. Sammon, "An optimal set of discriminant vectors,' IEEE Trans. Comput., vol. C-24, pp. 281-289, 1975.
[16] H. M. Kalayeh, M. J. Muasher, and D. A. Landgrebe, "Feature selection with limited training samples," IEEE Trans. Geosci. Remote Sen., vol. GE-21, pp. 434-438, 1983.
[17] K. Fukunaga and T. F. Krile, "Calculation of Bayes' recognition error for two multivariate gaussian distributions," IEEE Trans. Comput., vol. C-18, pp. 220-229, 1969.
[18] S. J. Whitsitt and D. A. Landgrebe, Error estimation and separability measure in feature selection for multiclass pattern recognition, School of EE, Purdue University, TR-EE 77-34, 1977.
[19] C. F. Hutchinson, "Techniques for combining Landsat and ancillary data for digital classification improvement," Photogrammetric Engineering and Remote Sensing, vol.48, no.1, pp. 123-130, 1982.
[20] R. M. Hoffer and staff, "Computer-aided analysis of Skylab multispectral scanner data in mountainous terrain for land use, forestry, water resources and geological applications," LARS Information Note 121275, Laboratory for Applications of Remote Sensing, Purdue University, W. Lafayette, IN 47907, 1975.
[21] P. H. Swain, J. A. Richards and T. Lee, "Multisource data analysis in remote sensing and geographic information processing," Proceedings of the 11th International Symposium on Machine Processing of Remotely Sensed Data 1985, W. Lafayette, IN., pp. 211-217, June 1985.
[22] Byungyong Kim, "Hierarchical Classification in High Dimension, Numerous Class Cases, PhD Thesis, Purdue University, West Lafayette, Indiana, December 1989.