

1-1-1982

A Binary Tree Feature Selection Technique for Limited Training Sample Size

M. J. Muasher

D. A. Landgrebe

Follow this and additional works at: <http://docs.lib.purdue.edu/larstech>

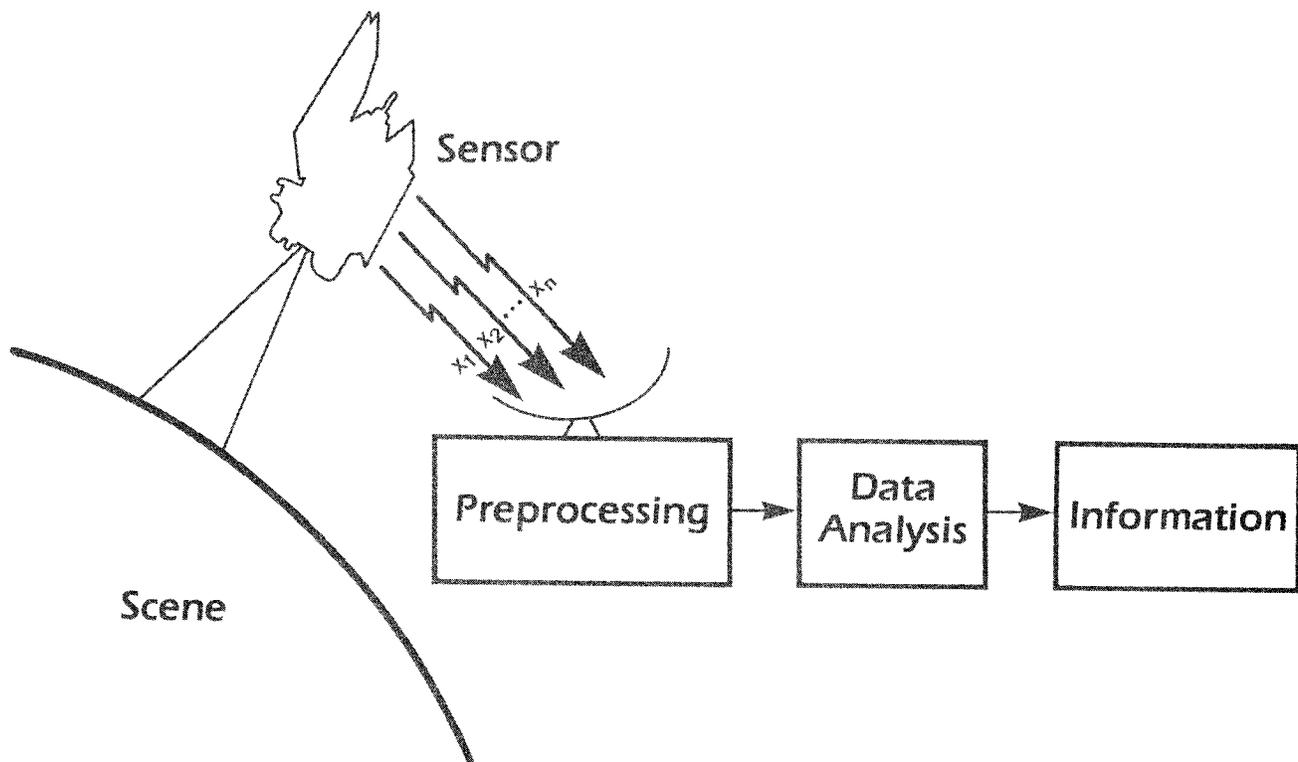
Muasher, M. J. and Landgrebe, D. A., "A Binary Tree Feature Selection Technique for Limited Training Sample Size" (1982). *LARS Technical Reports*. Paper 84.

<http://docs.lib.purdue.edu/larstech/84>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

Machine Processing of Remotely Sensed Data

with special emphasis on
Crop Inventory and Monitoring



July 7-9, 1982

Reprints from
Proceedings

Purdue University

**Laboratory for Applications of Remote Sensing
West Lafayette, Indiana 47907 USA**

A BINARY TREE FEATURE SELECTION TECHNIQUE FOR LIMITED TRAINING SAMPLE SIZE

M.J. MUASHER
D.A. LANDGREBE

Purdue University/Laboratory for
Applications of Remote Sensing

ABSTRACT

An algorithm is presented that predicts the mean recognition accuracy as a function of dimensionality for two-class problems, using a Bayes classifier in the presence of a limited number of training samples. Several experiments are presented to assess the algorithm's performance, and a binary tree classification procedure that utilizes the algorithm is shown to prove its usefulness.

I. INTRODUCTION

A number of different types of classifiers are now in use in remote sensing. Most of these classification techniques can be regarded as "single-stage" classifiers, where an unknown pattern is tested against all classes using one feature subset, and then the pattern is assigned to one of the present classes in a single stage classification procedure.

In recent years, the need has been felt for alternate, more powerful techniques through the use of which more information could be extracted from the scene. This is particularly important in the presence of a limited set of training samples because of the following reasons:

1. A characteristic of remote sensing problems is that training sample numbers are limited since in the remote sensing situation pre-labeled samples are usually difficult or expensive to obtain.
2. Current sensors in remote sensing applications produce small feature sets (usually 4) and limited gray scales and thus do not

require a large number of training samples to estimate class statistics. The next generation of sensors, beginning with the anticipated launching of the Thematic Mapper sensor, will produce larger feature sets, and more detailed gray scales, and hence will require larger numbers of training samples to adequately estimate class statistics.

3. Classification accuracy is known to be dependent upon feature set size, but current feature selection algorithms do not provide the ability to accurately determine at what subset dimensionality the best accuracy occurs.

In cases where there are larger numbers of features available than what should be used, current practice is to arbitrarily pick the number of features to be used, then to use a feature selection algorithm to determine the specific subset. We propose an algorithm that determines the optimal dimensionality, and the specific subset of features to be used, especially in the presence of the "Hughes Phenomenon"¹. Briefly stated, this phenomenon shows that in the presence of a limited training sample size, contrary to intuition, the mean accuracy does not always increase with additional measurements. Rather, it exhibits a peaking effect. Further, as the number of training samples increases, the peak occurs at a larger dimensionality, disappearing only in the case of an infinite number of training samples (complete knowledge of the underlying distributions). Any effective feature selection technique should be able to predict when/if this phenomenon occurs.

This paper presents a feature selection algorithm that takes into account the number of training samples used in estimating class statistics, then illustrates its use in a binary tree classification procedure, predicting the best feature subset to be used at each node. The procedure is particularly useful in cases where the Hughes Phenomenon occurs, as it is able to predict when the peak occurs and what feature subset to use in such cases. But it may also be useful where the number of training samples is large, predicting beyond what dimensionality accuracy increases would be so slight as to no longer be worth the added cost.

II. FEATURE SELECTION ALGORITHM

As mentioned earlier, our goal is to develop a performance estimator (to be used as a feature selection technique) that can predict the optimal subset of features. Some of the most serious difficulties facing researchers in trying to devise algorithms to estimate the probability of error in multidimensional analysis are:

1. Working with several features, the calculation of the probability of error requires an integration of a multivariate probability density function. Most often, this integration is almost impossible to carry out analytically, and very costly multivariate numerical integration has to be performed. Indirect methods which do not have a one-to-one relationship with the probability of error have been commonly used (Divergence, Bhattacharyya distance, ...etc.).
2. The measurement features are often correlated, making it difficult to assess the importance of each feature separately on the probability of error. Thus, all possible sets of features have to be compared, forcing the use of either very expensive calculations or sub-optimal techniques.
3. In most of the cases, one has to deal with multi-class problems (greater than two) which further complicates the integration on multivariate probability density functions. Also, in general, in multiclass cases the relationship between class pair error

rates and the overall rate is not one to one.

We seek a function that is one-dimensional, regardless of the number of features used. This will allow us to deal with a one-dimensional integration to calculate the probability of error, thereby reducing the complexity of the probability density functions. Such a function should retain all the information regarding the probability of error, which is what we are trying to estimate.

Fortunately, in the two class case, such a function does exist, and is called the likelihood function, defined as:

$$h(X) = - \ln p(X/w_1) / p(X/w_2) \quad (1)$$

where

$p(X/w_i)$ is the probability density function¹ of X given class w_i .

Assuming that $p(X/w_i)$ is multivariate normal, Fukunaga¹ and Krile² developed an algorithm which estimates the probability of error for multidimensional, two-class problems. However, their algorithm assumes accurate knowledge of the underlying distributions, and hence the probability of error they predict is monotonically decreasing with increasing dimensionality.

Muasher and Landgrebe^{3,4} modified Fukunaga and Kriles' algorithm by taking into account the number of training samples used in estimating the statistics of the two classes at hand. The probability of error, P_E , which is the area of overlap under the probability density functions of h/w_1 and h/w_2 (multiplied by the prior probabilities) can change considerably if the estimated parameters of h/w_i are poor as a result of an inadequate training sample size. The algorithm developed in (3,4) looks at the variances of h/w_1 and h/w_2 , σ_1^2 and σ_2^2 , and computes their variances. It shows that as the number of features increases, the variances of σ_1^2 and σ_2^2 increase rapidly, offsetting the increase in separability between classes and thus leading to a peaking effect.

A new, modified algorithm is then developed (See (3) for complete details) to take into account the number of training samples. The algorithm estimates the probability of error by approximating the area under the likelihood ratio function for two classes, taking into account the number of training samples used in estimating each of these two classes. In the next section, results are

presented that compare the performance of the algorithm against experimental observations. Also, a binary tree classification which uses the algorithm for feature selection is shown to illustrate the usefulness of the procedure.

III. RESULTS

Two data sets are used in our experiments: An aircraft data set, and a Landsat set. The aircraft data set was collected on August 13, 1971, over Tippecanoe County, Indiana, and has 12 spectral bands. The Landsat set was collected over Henry County, Indiana. A multitemporal data set was constructed by registering four data sets flown over the site at different times. The dates the data were collected on are: June 9, July 16, August 20, and September 26, all in 1978. It was established in (3,4) that the Karhunen-Loeve ordering method, in which the features are ordered according to descending eigenvalues after a K-L transformation is performed on the data set, is an effective feature selection technique in the presence of a limited number of training samples. This method will be used here, and consequently, a K-L transformation was performed on both data sets; the first 12 channels in each set were used for classification.

Both real and simulated data are used. The simulated data is based on the statistics of real data, using a method described in (5). The purpose of simulating data is to satisfy several assumptions that are commonly made in remote sensing, but not always exactly satisfied with real data. These assumptions include class-conditional multivariate normal distributions, known number of classes in the scene, and "pure" pixel elements. The simulation technique used preserves the natural spatial information occurring in multispectral data by spatially basing the simulation on a classification map.

Two classes are used in each data set: Corn and forest in the aircraft data set, and corn and soybeans in the Landsat set. In each case, a large number of samples per class is chosen for training, and a larger, mutually exclusive set is used for testing. Five training sets, each one having 20 samples per class, are randomly chosen from each of the larger training sets. Another 5 training sets are also chosen, but with each set having 13 samples per class, the minimum number one can use without

a singular covariance matrix resulting in 12 dimensions. The K-L method is used for ordering the features, and the test fields are classified, using the statistics obtained from the 5 training sets. The average classification accuracy, P_{CC} , over the 5 sets, is calculated for the best 2,3,...,12 feature subsets. Also, the results obtained by using the proposed algorithm are plotted versus experimental observations to compare the two.

Results appear in Figure 1 for aircraft data, and Figure 2 for Landsat data. Also plotted are the standard deviations of errors for each feature subset using the 5 different training sets.

Results indicate that the algorithm predicts the best, or near best, subset of features to be used. The algorithm results have the same shape as the trends in the corresponding experimental curves.

The algorithm also predicts the P_{CC} values within a few percent. Since the objective behind the algorithm is to predict the best feature dimensionality and specific feature subset to be used in classification rather than to predict the probability of error itself, the fact that the algorithm does not always predict this probability of error with an arbitrarily small prediction error is not of concern. It is worth noting here that the plotted experimental curves are averaged plots of random variables (i.e. the result of several trials of a random experiment) while the algorithm result is an average value (expected value) and therefore not a random variable.

The standard deviations plotted tend to confirm the expected trend that in general, an increase in dimensionality results in an increase in the variance of error, that increase becoming highly noticeable at high dimensionality, when the randomness in the estimated statistics given a fixed, finite set of training samples, is large. This is further confirmed in Muasher and Landgrebe⁴.

The next step is to incorporate this algorithm in a binary tree classification procedure. The aircraft data set is used here. Nine spectral classes exist in the scene. 13 samples per class are used for training, with a larger, mutually exclusive set for testing. The binary tree is constructed by using a bottom-up procedure, combining the two most separable classes each time, and using a separability measure developed by Whitsitt and Landgrebe⁶, and defined as follows:

$$D_{\text{erf}} = \text{erf}(\sqrt{2B}) \quad (2)$$

where B is the Bhattacharyya distance and erf(.) is the Gaussian error function.

The proposed algorithm is used to predict the optimal features at each node. A single-stage, maximum likelihood classification is then performed on the two sets, using feature subsets of 2 to 12. This is done to compare the performance of the binary tree procedure to that of the feature subsets.

Results are shown in Figure 3. The numbers below each node indicate the features used at these nodes, and the numbers inside the nodes indicate the number of training samples at each node. The figure below the tree shows the results of the single-stage and the binary tree classifiers.

Results indicate that the algorithm is effective in predicting the feature subsets that lead to the maximum accuracy possible using the K-L transformation for ordering the features. In the example, the binary tree procedure results in the maximum (or even better) accuracy possible using a single-stage classifier, but with the added advantage that it provides a method for selecting those feature subsets which lead to the maximum accuracy.

It is worthwhile to note that common belief has been that fewer features need be used at the top of the tree to separate classes, and more features need be used deeper in the tree to distinguish between somewhat inseparable classes. However, if there are inadequate training samples present, then the number of training samples towards the bottom of the tree is less than that towards the top. Hence, less features should be used at the bottom to avoid a peaking effect.

IV. CONCLUSION

The proposed algorithm for feature selection appears to be effective in predicting the best feature subsets to use in the presence of a limited number of training samples. The algorithm is especially useful in a binary tree classification procedure, where it is shown to predict the best accuracy possible in a fairly involved data set (9 classes, 12 features).

The program provides the ability to use so small a number of training samples and still get the best classification

accuracy possible out of the available statistics. Moreover, results seem to indicate that the rule of thumb often used in remote sensing applications, stating that the number of training samples should be 10 times larger than the number of features used, might be too high. Indeed, working with such small numbers of training samples in multispectral data is new.

V. REFERENCES

1. Hughes, G.F. (1968). On the Mean Accuracy of Statistical Pattern Recognizer. *IEEE Trans. Infor. Theory*. IT-14(1):55-63.
2. Fukunaga, K. and T. Krile. (1969). Calculation of Bayes Recognition Error for Two Multivariate Gaussian Distributions. *IEEE Trans. Computers*. C-18(3):220-229.
3. Muasher, M.J. and D.A. Landgrebe. (1981). Multistage Classification of Multispectral Earth Observational Data: The Design Approach. 171p. Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana. LARS Technical Report 101481. Also available as a Ph.D Thesis, TR-EE-81-41. School of Electrical Engineering, Purdue University.
4. Muasher, M.J. and D.A. Landgrebe. (1982). On the Hughes Phenomenon: Some Experimental Observations and Theoretical Derivations. To be published.
5. Muasher, M.J. and P.H. Swain. (1980). A Multispectral Data Simulation Technique. 30p. Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana. LARS Technical Report 070980.
6. Whitsitt, S.J. and D.A. Landgrebe. (1977). Error Estimation and Separability Measures in Feature Selection for Multiclass Pattern Recognition. 186p. Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana. LARS Publication 082377. Also available as a Ph.D Thesis, TR-EE-77-34. School of Electrical Engineering, Purdue University.

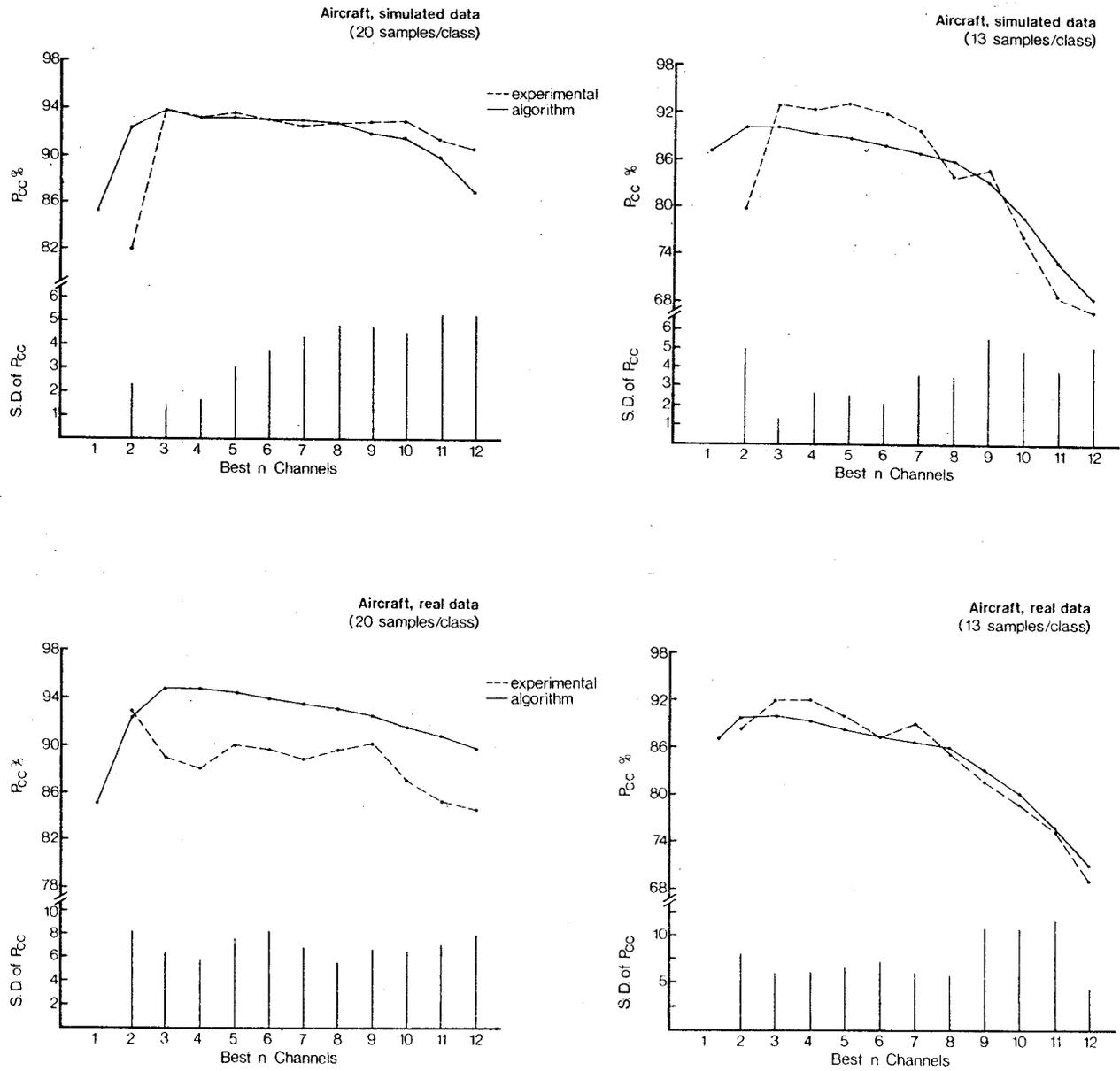


Figure 1. Experimental and algorithm results for aircraft data.

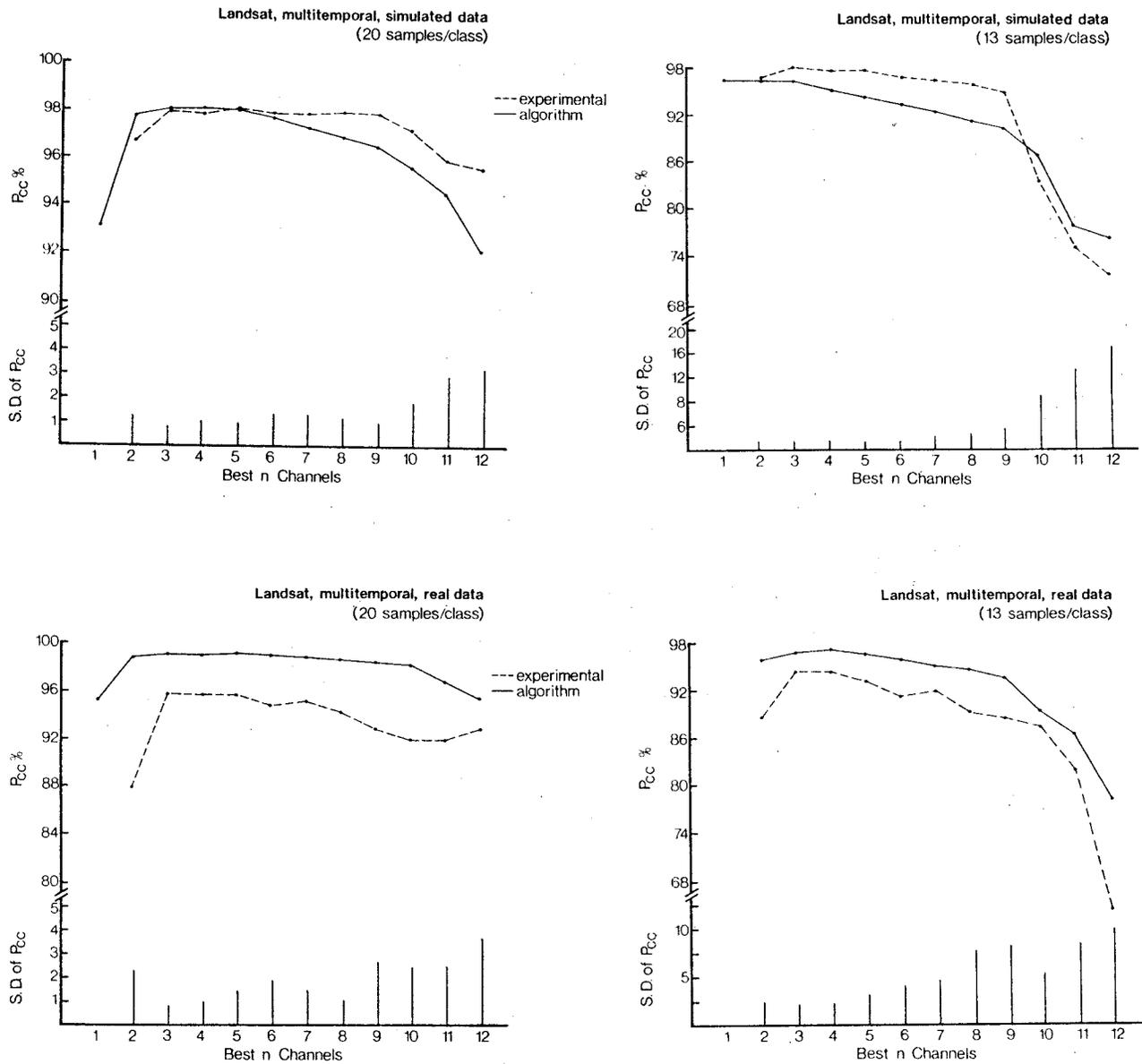


Figure 2. Experimental and algorithm results for Landsat data.

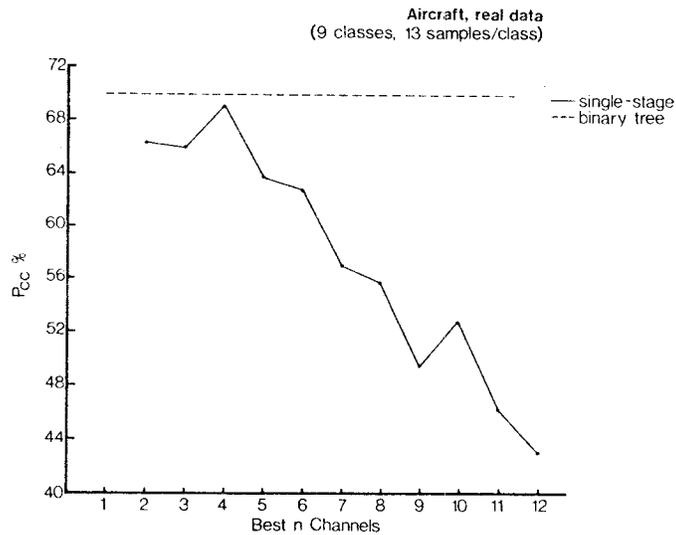
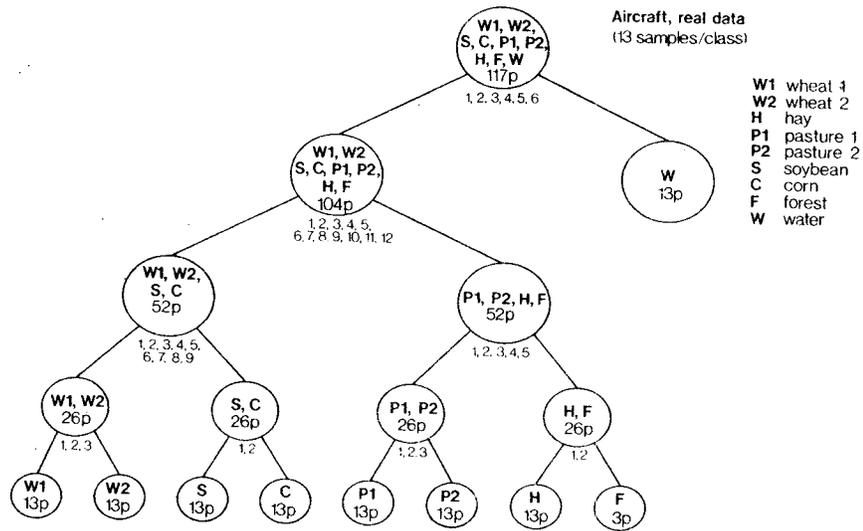


Figure 3. Single-stage and binary tree classification results for the aircraft data set.

Marwan Jamil Muasher was born on June 14, 1956 in Amman, Jordan. He received his B.S.E.E from Purdue University in 1977. He enrolled in the graduate school at Purdue and received his M.S.E.E and Ph.D degrees in 1978 and 1981 respectively. From January 1978 to October 1981 he was employed as a research assistant at the Laboratory for Applications of Remote Sensing. Dr. Muasher is a member of IEEE, Tau Beta Pi, and Eta Kappa Nu.

David Allen Landgrebe is the Associate Dean of Engineering at Purdue University, and Director of the Engineering Experiment Station. He holds the B.S.E.E, M.S.E.E, and Ph.D from Purdue. He joined the Purdue EE faculty in 1962. He was named Program Leader for Data Processing Program at Purdue's LARS in 1966, and served as its Director from 1969 until 1981. He received the NASA Exceptional Scientific Achievement Medal in 1973. Dr. Landgrebe is a fellow of IEEE, and a member of several professional and honorary organizations. He is also an associate editor of the journal, Remote Sensing of the Environment, and a member of the administrative committee of the IEEE Geoscience and Remote Sensing Society.