LARS Information Note 061275

LAYERED CLASSIFICATION TECHNIQUES FOR REMOTE SENSING APPLICATIONS

BY

P. H. SWAIN
C. L. WU
D. A. LANDGREBE
H. HAUSKA

The Laboratory for Applications of Remote Sensing

Purdue University, West Lafayette, Indiana

1975

LARS Information Note 061275

LAYERED CLASSIFICATION TECHNIQUES FOR REMOTE SENSING APPLICATIONS+

P. H. Swain, C. L. Wu, D. A. Landgrebe, and H. Hauska
Laboratory for Applications of Remote Sensing,
Purdue University, West Lafayette, Indiana*

ABSTRACT

Layered classification offers several advantages over the very familiar single-stage approach. The single-stage method of pattern classification utilizes all available features in a single test which assigns the "unknown" to a category according to a specific decision strategy (such as the maximum likelihood strategy). The layered classifier classifies the "unknown" through a sequence of tests, each of which may be dependent on the outcome of previous tests. Although the layered classifier was originally investigated as a means of improving classification accuracy and efficiency, it has become apparent that in the context of remote sensing data analysis, other advantages also accrue due to many of the special characteristics of both the data and the applications pursued. This paper outlines the layered classifier method and discusses several of the diverse applications to which this approach is well suited.

INTRODUCTION

The pattern classifiers commonly implemented in remote sensing data processing systems have an elementary logical structure like that shown in Figure 1. To classify each "unknown," the decision process is applied to a fixed set of features (usually spectral measurements) to select one of the candidate pattern classes. For some remote sensing problems encountered in practice, however, this simple structure is less than ideal and may even be inadequate. A structure like that suggested by Figure 2, a "decision tree" or "layered classifier," can be put to good use. The following cases illustrate this fact.

In regions characterized by highly variable topography, spectral response may depend significantly on topographic variables such as slope and aspect [1]. With the digital data bases and the data registration facilities which are now available, it is possible to overlay, say, satellite multispectral scanner (MSS) data and topographic data. Thus, it is feasible to use the topographic data in the classification process - except that the topographic data cannot be handled as simply an additional variable (or variables). Ideally, one might wish to use the topographic variable(s) to stratify the multispectral data analysis problem, and then apply a multivariate maximum likelihood procedure to the stratified data. This can be accomplished with a layered decision procedure, in which the data is stratified in the first layer and the classification is accomplished in one or more succeeding layers.

---

*Dr. Wu is now with the Jet Propulsion Laboratory, Pasadena, California; Dr. Hauska is an ESRO Research Fellow, at Purdue University on leave from Luleå University of Technology, Luleå, Sweden.

+Presented at the Earth Resources Survey Symposium, Houston, Texas, June 9, 1975.

Another case in which a kind of stratification is of use is in detection of diseased crops. The spectral bands which are optimal for disease detection are not necessarily the bands optimal for crop species discrimination. Rather than use all of these bands jointly for every classification, which may be prohibitively expensive in terms of computational load, it would be preferable to use the "disease detection bands" only when the crop species of concern has been detected. Again, a multilevel decision procedure can be used.

Another class of problems for which layered classification is helpful is that characterized by very limited size training sets. Under such limitations, it may be necessary to restrict the dimensionality (number of features) of any decision procedure used in order to avoid the "dimensionality problem;" i.e., the phenomenon which, when the number of training samples is small, decreases classification accuracy when new pattern features are added [2,6]. This problem can often be circumvented by substituting multiple decisions with low dimensionality for a single decision having high dimensionality, which can be accomplished using layered decision logic.

Many other examples could be cited. In general they involve one or more of three factors: (1) compatibility of the scene variables with each other and with the analysis process; (2) the desire to maximize classification accuracy; (3) the desire to minimize classification cost (usually measured in terms of computer time). The challenge is to devise a systematic and effective procedure for determining the optimal decision logic.

APPROACH

For generality, we shall state the problem as follows:

Assume there is a set of classes $C = \{c_1, c_2, \ldots c_m\}$, a set of measurements or features $F = \{f_1, f_2, \ldots, f_n\}$ with which to discriminate between the classes, and a set of decision procedures $P = \{p_1, p_2, \ldots, p_k\}$ which can be applied to the features to achieve the required discriminations. Formally, a "decision tree" is a tree-like structure of connected nodes, such as shown in Figure 3, which defines a logical decision (classification) procedure. Each node (junction point) in the tree is labelled by a triple of the form $\{C_i, F_i, p_i\}$, where C_i is a subset of C, F_i is a subset of F, and P_i is a single element from P. Thus each triple defines a step in the decision process for classifying any given data observation. The uppermost or root node in the tree is always labelled with C; the lowest or terminal nodes in the tree are often, but not always, labelled with individual elements of C.

The problem is to select from the set of all possible decision trees a tree which (1) at every node uses a decision procedure compatible with the available data features, and (2) maximizes the overall classification accuracy while (3) minimizing the classification cost. Conditions (2) and (3) must be met in terms of an "optimal" trade-off. For almost any practical problem, the specifications for some of the nodes in the tree will follow naturally from the problem itself. But there may be a very large number of possibilities for the remainder of the nodes. Designing an "optimal" decision tree requires first a criterion for evaluating any given tree, and then a strategy for searching for the optimal tree among the possibilities. In the remainder

of this section we shall describe briefly a design approach we have been developing. The details may be found elsewhere [2].

Optimality Criterion

For a given data processing system, the computation time required to analyze the data of interest reflects the relative "cost" of the analysis. For instance, the time increases with the complexity of the decision procedure and the number of pattern features used. On the other hand, it is usually possible, up to a point, to "buy" analysis accuracy by paying the cost of a more complex decision procedure and/or adding more features.

The optimality of a given decision tree is expressed in terms of both the time required by the classifier to process the data and the error rate achieved. Since, up to a point, one of these factors can be traded for the other, the relative importance of these factors must be specified for the problem at hand. The optimality criterion is expressed as an "evaluation function" E:

$$E = -(T + K \cdot \varepsilon) \tag{1}$$

where T is the time required for classification and ε is the resulting error rate. The constant K expresses the relative importance of computation cost (time) and classification error. We have chosen to use the negative expression so that achieving optimality corresponds to maximizing the evaluation function.

Searching for an Optimal Tree

For any given problem, it would be theoretically possible to enumerate (make an ordered exhaustive list of) all possible decision trees and evaluate them one by one using eq. (1) to find the best possible tree. But for almost any nontrivial problem encountered in practice, this is infeasible because of the enormous number of possible trees. To get this problem down to manageable proportions, a strategy has been adopted which we refer to as "guided search with forward pruning" [2]. Essentially, the strategy is to construct the tree a node at a time, estimating the suitability of all candidate structures for the node under consideration, and discarding all but the most promising candidate "subtree."

The "guided search with forward pruning" strategy requires a means for evaluating each node. For each candidate structure following node d_i, the evaluation is computed as follows:

$$E(d_i) = -T(d_i) - K \cdot \varepsilon(d_i) + \sum_{j=1}^{n_i} E(d_{i+j}) \tag{2}$$

where $T(d_i)$ and $\varepsilon(d_i)$ express the efficiency and accuracy of the node d_i and the summation is an estimate of the evaluation functions of the descendant nodes of d_i (which are n_i in number). A lower bound is used to form this estimate, which is the evaluation of a conventional single-stage classifier applied at that point.

Constructing the decision tree in this sequential fashion cannot guarantee that the optimal tree will be obtained, because, unfortunately, the optimal choice at any level in the tree is not necessarily independent of choices at later stages. However, improvement over conventional single-stage classifiers is generally achievable, a fact amply demonstrated by the empirical results obtained to date [2,3].

Estimation of Classification Accuracy

Use of eq. (2) requires computation of classifier error in order to evaluate any candidate subtree. However, even under the simplifying assumption that all classes to be recognized have multivariate Gaussian statistics, direct calculation of classifier error is not feasible [4]. Instead it is necessary to use an indicator function which provides a measure of the statistical separability of the pattern classes for any given subset of the pattern features. Several such functions have been investigated for this purpose, a transformation of the Bhattacharyya distance having proved most suitable so far [2,3,4].

Alternative Decision Tree Design Procedures

The search strategy discussed above provides an analytical tool for designing decision trees based on the statistical relationships between the pattern classes. In some instances, however, the appropriate design criteria may not be of a statistical nature. In such cases, a more user-interactive, less automatic procedure may be followed. Typically, the user will specify those nodes of the decision tree which can be specified from consideration of the problem characteristics and then use the optimal search method to determine the remainder of the tree. A good example is the approach for overcoming cloud problems, which is described in the next section.

APPLICATIONS

In the Introduction, we cited two potential applications and an entire class of problems for which layered classifiers are useful. Several more applications are listed in Table I. We will now describe a number of instances where this method has actually been used experimentally.

Water Temperature Mapping

It was desired to map the surface temperature of a river for which a multispectral scanner data set was available including a calibrated thermal channel [5]. To accomplish the desired mapping, it was first necessary to locate the river using the visible and reflective infrared scanner channels. The temperature of the water was then determined from the thermal infrared channel.

Figure 4 shows a simple decision tree structure used to accomplish the desired result. The design was simple enough to perform manually, without the use of the optimal search procedure.

Avoiding the Clouds

The classification of agricultural crops from satellite multispectral data promises to be one of the major applications of remote sensing. However, at many of the latitudes where important cropland is found, the remote sensing

data is often cloudy. A layered classifier used in conjunction with data from multiple satellite passes over the same area can alleviate this problem to a substantial degree.

In an experiment which demonstrated this capability, LANDSAT data collected over Livingston County, Illinois on June 29 and July 16, 1973, were geometrically registered. It was desired to determine, using the data from the June 29 pass, the acreage of corn and soybeans planted in the county. However, ten percent of the county was obscured by clouds and cloud shadows. Training statistics were determined for both dates, and the decision logic shown in Figure 5 was used to classify the data. Again, as in the previous case, the decision tree was determined by the problem logic, without resorting to the optimal search procedure.

As shown in Figure 5, the layered classifier was used simply to eliminate the effects of clouds in the earlier data set. More complex logic could have been developed. For example, at points where neither data set had clouds or shadows, all eight data channels could have been used for classification with the aim of improving the classification accuracy.

Optimization of Accuracy and Efficiency

A number of applications examples are described in [2] which involve designing layered classifiers so as to optimize accuracy and efficiency. In one fairly typical case it was desired to identify the classes "coniferous forest," "deciduous forest," "agriculture," "water," and "bare rock." In the analysis process, twenty-six spectrally distinct subclasses were identified, and the optimal search procedure was used to derive an appropriate decision tree, which is shown in Figure 6. (A simplified notation has been used in this figure to specify the tree. All nodes use maximum likelihood as the decision criterion.) This tree is fairly typical in several respects, including its breadth and depth. Efficiency of the classification process is gained through the fact that many of the decisions involve a small number of features (for the classification algorithm used, the time required is roughly proportional to the square of the number of features). Accuracy is gained by using an optimal subset of features for each decision, which is generally a subset determined by the specific classes to be discriminated in that decision.

SUMMARY

We have described here a flexible method for classification of remote sensing data. This method can be applied to a wide range of problems and varieties of data not conveniently handled by conventional classifiers. It also provides a means of maximizing classification accuracy and efficiency by optimizing the number and selection of features used in each decision.

The "decision tree" approach allows the user to take advantage of special characteristics of the problem at hand by manually specifying parts of the decision tree. Where the problem is one of choosing among a vast number of alternative tree structures, a programmed "optimal search procedure" is available. Often a hybrid of the manual and optimal search procedures provides the most suitable decision tree structure for a practical problem.

For problems involving large numbers of pattern classes and features, the "optimal search procedure" described here cannot guarantee that the resulting decision tree is truly the optimal tree. Future research using

heuristic search procedures and mathematical programming techniques is expected to improve on the design tools now available.

ACKNOWLEDGEMENT

The research reported in this paper was supported by NASA Grant NGL 15-005-112 and NASA Contract No. NAS 9-14016.

REFERENCES

1. R. M. Hoffer, M. D. Flemming, and P. V. Krebs, "Use of Computer-Aided Analysis Techniques for Cover Type Mapping in Areas of Mountainous Terrain," Information Note 091274, Laboratory for Applications of Remote Sensing (LARS), Purdue University, West Lafayette, Indiana 47907; September 1974.

2. C. L. Wu, D. A. Landgrebe, and P. H. Swain, "The Decision Tree Approach to Classification," Information Note 090174, Laboratory for Applications of Remote Sensing (LARS), Purdue University, West Lafayette, Indiana 47907; May 1975.

3. H. Hauska and P. H. Swain, "The Decision Tree Classifier: Design and Potential," Proceedings Purdue Symposium on Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana, June 1975.

4. P. H. Swain and R. C. King, "Two Effective Feature Selection Criteria for Multispectral Remote Sensing," Proceedings International Joint Conference on Pattern Recognition, Washington, D. C., November 1973.

5. L. A. Bartolucci, R. M. Hoffer, and T. R. West, "Automatic Data Processing of Remotely Sensed Data for Temperature Mapping of Surface Water," Information Note 042373, Laboratory for Applications of Remote Sensing (LARS), Purdue University, West Lafayette, Indiana 47907; August 1973.

6. G. F. Hughes, "On the Mean Accuracy of Statistical Pattern Recognizers," IEEE Transactions on Information Theory, volume IT-14, no. 1, pp. 55-63, January, 1968.

Table I. SOME APPLICATIONS OF LAYERED CLASSIFIERS

General Application	Example
Change Detection	Snow pack variation Water level variation (e.g., reservoirs) "Urban sprawl" Logging practices
Use of Mixed Feature Types	Texture Topography Geophysical data (e.g., aeromagnetic)
Class-specific Properties	Crop disease detection Forest type mapping Water quality mapping Water temperature mapping (see text)
Other	Avoidance of cloud effects (see text) Minimization of data dimensionality

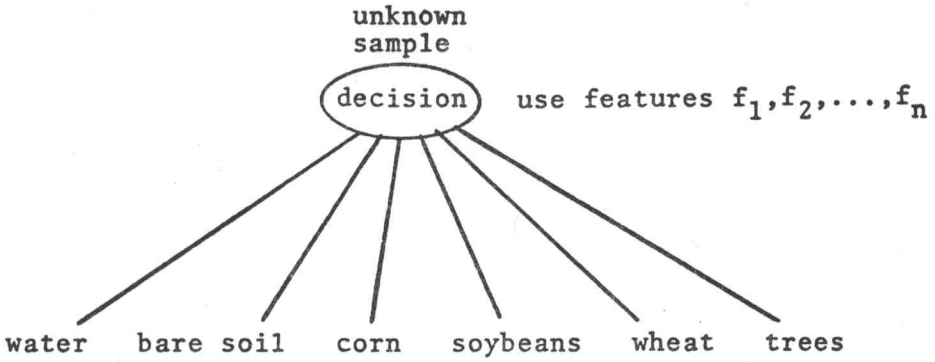


Figure 1.- Common single-stage classifier.

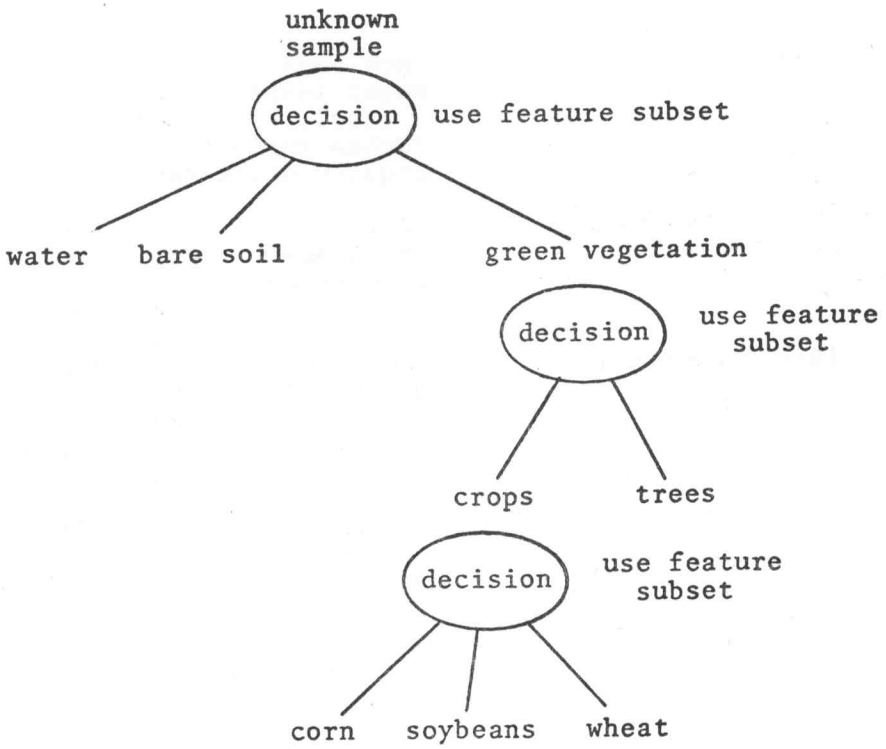


Figure 2.- A layered classifier.

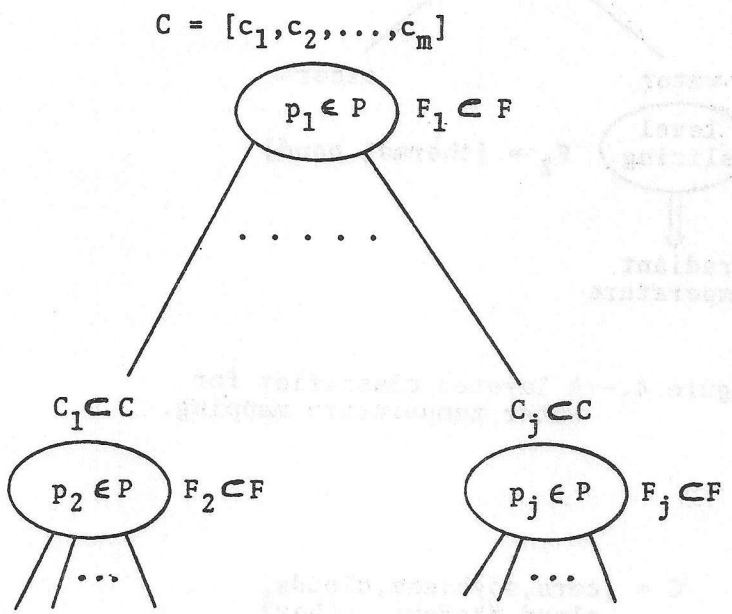


Figure 3.- General form of the decision tree.

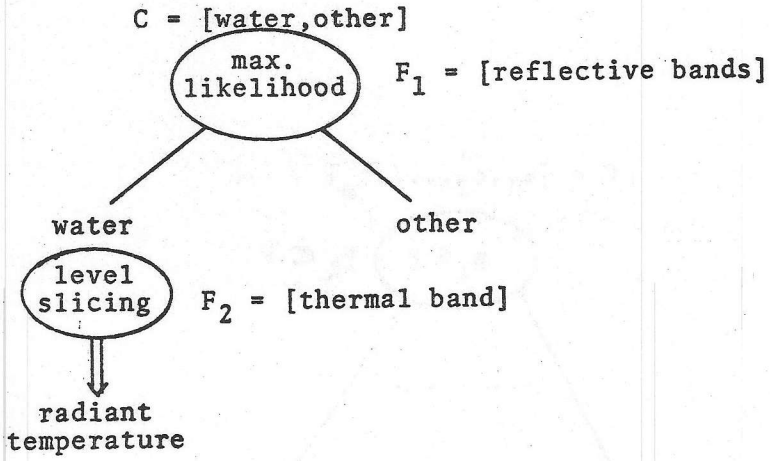


Figure 4.- A layered classifier for water temperature mapping.

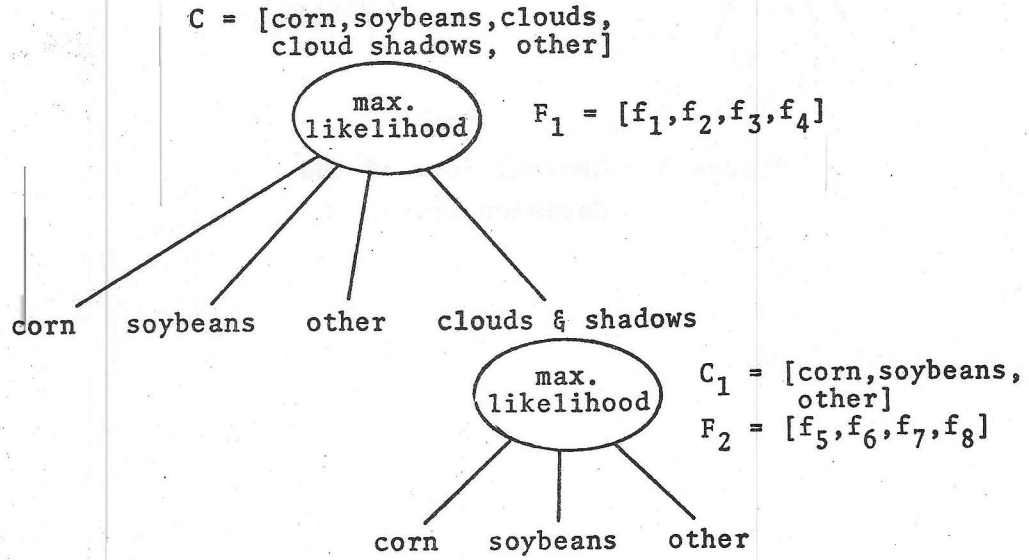
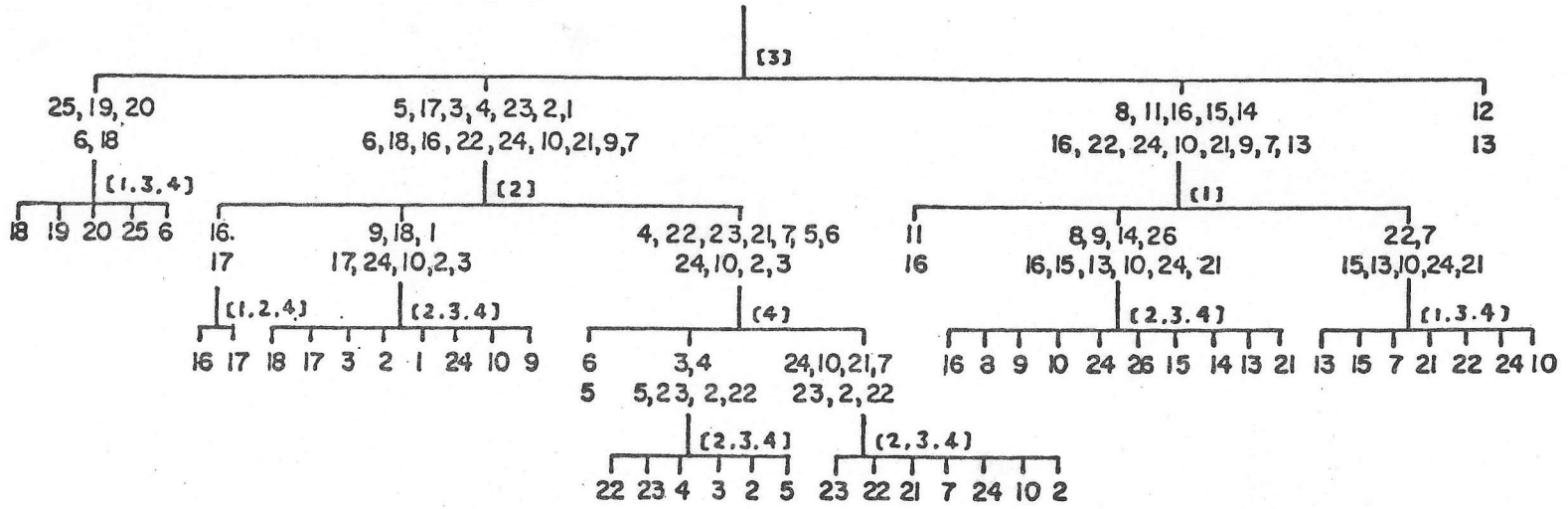


Figure 5.- Agricultural classification in the presence of clouds.



1. All decisions are maximum likelihood.

2. [ ] indicates features used.

3. Conifer: classes 1,2,3,4,5,6,23,25.
   Deciduous: classes 7,8,9,10,24,26.
   Agriculture: classes 11,12,13,14,15.
   Bare rock: class 17.
   Water: classes 18,19,20.

Figure 6. A layered classifier designed for accuracy and efficiency.