C. Gleason 11/74

LARS Information Note 041574

AN AUTOMATED AND REPEATABLE DATA ANALYSIS PROCEDURE FOR REMOTE SENSING APPLICATIONS

B. J. DAVIS
P. H. SWAIN

The Laboratory for Applications of Remote Sensing

Purdue University, West Lafayette, Indiana

1974

LARS Information Note 041574

AN AUTOMATED AND REPEATABLE DATA ANALYSIS PROCEDURE FOR REMOTE SENSING APPLICATIONS*

B. J. Davis and P. H. Swain

Laboratory for Applications of Remote Sensing
Purdue University
West Lafayette, Indiana 47907

ABSTRACT

A new multispectral data analysis procedure, based on LARSYS, has been developed which substantially reduces the influence of the analyst. The analysis is automated, including the interpretation of clustering results. The classification results obtained are repeatable and not biased by analyst subjectivity during the analysis.

1. INTRODUCTION

At present several different data processing systems and procedures are available for the analysis of multispectral remote sensing data. These vary greatly in the extent to which the human analyst plays a role in the analysis; but in all cases the analysis results are at least in some measure influenced by the knowledge, skill, and experience of the analyst, by his interpretation of both the data set and the analysis procedure, and by the question he is trying to answer in carrying out the analysis. This can be advantageous: the analyst can tailor the procedure to his particular problem and data set with considerable benefit. But at the same time this dependency may prevent extension of the results to other problems, data sets, and analysts. It may also preclude objective comparisons of analyses of different data sets unless the influence of the analyst(s) can be very carefully quantified and/or controlled. The variation among analyses due to the analyst becomes a particularly serious issue in remote sensing research intended to evaluate different data collection or data processing technologies.

And finally, of course, dependence of human intervention in the analysis process implies a sacrifice in speed over methods which are completely automatic, which certainly impacts future applications requiring high-speed high-volume processing.

2. A NEW ANALYSIS PROCEDURE

A new multispectral data analysis procedure, based on LARSYS, has been developed which substantially reduces the influence of the analyst. By eliminating all intermediate interpretation steps involving an analyst, it provides an unbiased, repeatable classification procedure and decreases the time required.

The new data analysis procedure has four major steps:

1) specification of the problem to be analyzed,
2) clustering of the training sets to isolate the spectral subclasses with unimodal, approximately Gaussian distributions which are present in the classes specified by the defined problem,
3) interpretation of the cluster results, and
4) classification of the data set into the classes defined in the problem specification.

*The research reported herein is supported by NASA Grant NGL 15-005-112.

The first step -- the specification of the problem -- is the stage at which human choices are made. Once the decisions involved in specifying the problem are made, the remaining steps of the procedure carry out these decisions automatically.

2.1. SPECIFICATION OF THE ANALYSIS PROBLEM

The specification of the problem to be analyzed involves four decisions for the analyst. The analyst must choose the classes which are the subject of the analysis. He must decide how many features or wavelength bands are to be used in the classification of the data set. The analyst must specify the training set for each class. Finally, the number of clusters into which the training set is to be clustered must be specified for each class.

There are numerous ways in which the number of clusters for each class can be chosen. One can merely use an estimate of the number of spectrally distinct subclasses in each class; another possibility is to base the number of clusters on a function of the number of data points in the training set.

Another possible procedure for deciding the number of clusters to be used is the following rule. It makes use of an estimate of the distinct spectral subclasses in a class and of the number of data points in the class's training set to help insure that the statistics representing the subclasses are valid. The analyst must already have defined the classes in the analysis: $C_1,\ldots,C_m$, and decided upon n, the number of features to be used in the classification. The analyst must have also specified the associated training sets: $T_1,\ldots,T_m$ which contain $t_1,\ldots,t_m$ data points, respectively. An estimate of the number of spectrally distinct subclasses is also needed: $S_1,\ldots,S_m$. From the number of training data points, t_i, and from the estimate of the number of spectrally distinct subclasses, S_i, the number of clusters to be considered for each class, ξ_i, can be determined.

Step 1. Set $\xi_i = 3S_i$ for $i = 1,\ldots,m$.

Step 2. If $\xi_i \geq [t_i/10n]$*, set $\xi_i = [t_i/10n]$. This insures that the number of points in each cluster will be approximately 10 times the number of features used for classification.

Step 3. If $\xi_i < S_i$, set $\xi_i = S_i$.

2.2. CLUSTERING OF TRAINING SETS

When the number of clusters to be used in each class is decided, the training set for each class is clustered into the specified number of clusters by the cluster processor. The LARSYS cluster processor implements an iterative, Euclidean-distance clustering algorithm [1]. Initial cluster centers are chosen and then all the training data points are assigned to the nearest cluster center. The cluster centers are replaced by the means of the current clusters and the points are again assigned to the nearest cluster center. The process continues until no data point changes its cluster "allegiance".

The results from the cluster processor include a measure of cluster separability for each pair of clusters. This separability measure, a distance quotient, is based on both the means and covariances of the clusters and compares the separation of the cluster centers to the within-cluster dispersion.

Figure 1 illustrates the derivation of the distance quotient [2]. Two clusters and their respective ellipsoids of concentration [3] are shown in Figure 1. D_{ij} is the distance between the cluster centers. D_i is the distance from the center of cluster i to the surface of its ellipsoid of concentration along the line connecting the cluster centers. Similarly, D_j is the distance from the center of cluster j to the surface of its ellipsoid of concentration along the line connecting the cluster centers. The distance quotient, d_{ij}, is given by:

$$d_{ij} = \frac{D_{ij}}{D_i + D_j}.$$

*[X] is the integer part of X.

Cluster i and j are considered distinct when $d_{ij} > T$ where T is a suitable threshold. The threshold generally used in this procedure is 0.75, as it has been observed empirically that a single class containing two clusters for which $d_{ij}>0.75$ will generally have a multimodal distribution.

2.3. INTERPRETATION OF CLUSTER RESULTS

The interpretation of the set of distance quotients which are defined for all pairs of clusters has been one of the most crucial and yet analyst-dependent steps in the analysis procedure. Heretofore, the analyst subjectively decided how clusters would be combined into subclasses. To eliminate the need for analyst interpretation and intervention, the following cluster separability interpretation algorithm has been programmed into the cluster processor and produces a table which indicates which clusters are spectrally similar and may be combined together into subclasses having unimodal, approximately Gaussian distributions.

2.3.1. CLUSTER INTERPRETATION ALGORITHM.

Step 1. Assign each cluster to its own cluster group, $C_1, C_2, \ldots, C_n$.

Step 2. Order the set of distance quotients previously defined, $\{d_{ij}\}$, by magnitude. The algorithm considers each d_{ij} in order, beginning with the smallest. Let d_{xy} equal the smallest such d_{ij}.

Step 3. If $d_{xy}>T$ (where T is the threshold of 0.75 mentioned previously), stop. Otherwise, proceed to step 4.

Step 4. If clusters x and y belong to the same cluster group ($C_x = C_y$) set d_{xy} to the next (larger) value of d_{ij}, and return to step 3. Otherwise, proceed to step 5.

Step 5. Construct the average distance $\bar{d}_{xu}$ between C_x and each other cluster group $C_x \neq C_u$ for which $d_{ab} \leq T$ for all clusters a in C_x and clusters b in C_u. The average distance between cluster groups is defined as the average of all pairwise distances between clusters in the different cluster groups.

Step 6. If $\bar{d}_{xy}$ is the minimum of the set of inter-group distances constructed in step 5, then combine C_x and C_y into one cluster group.

Step 7. Set d_{xy} to the next d_{ij} and return to step 3.

This algorithm provides a systematic procedure for interpreting the separability information. When the algorithm is finished, each cluster group, containing one or more clusters, represents a subclass of the class which was clustered. The procedure given above minimizes the total number of subclasses produced while ensuring that multimodal subclass distributions are avoided.

2.4. CLASSIFICATION OF THE DATA SET

After the clustering results are interpreted for all classes, individually, each subclass can be represented by a mean vector and covariance matrix associated with a cluster group (if the cluster group contains more than one cluster, the statistics for all clusters in the cluster group are statistically pooled). Each point in the data set is then classified by the maximum likelihood classification rule. The decision rule for classification which is implemented in LARSYS [1] is:

Classify X as belonging to class ω_i if

$$g_i(X) \geq g_j(X) \text{ for all } j \neq i$$

where X is the data vector and

$$g_i(X) = -1/2 \log |K_i| - 1/2 (X-M_i)^T K_i^{-1} (X-M_i)$$

M_i and K_i are, respectively, the mean vector and covariance matrix for class ω_i.

An estimate of the accuracy of the classification results can be made through the use of test fields and machine tabulation or by a manual comparison of the analysis results to known areas of interest.

3. SUMMARY AND CONCLUSIONS

In this procedure, the influence of the analyst has been limited to the first step of the procedure, the specification of the problem. In defining the problem, the analyst uses his knowledge and experience to make reasonable decisions in specifying the classes, features, training sets, and number of clusters. But once the four decisions are made, the analysis proceeds automatically, and the analyst does not influence the process. As long as the problem specification remains the same, the classifications results do not depend on the analyst and so are completely repeatable.

Since the analyst does not influence the analysis after the decisions are made in the problem specification, the classification results truly reflect the choices made in the specification of the problem, and so objective comparisons of the effects of different decisions are easier to make than in the conventional analyst dependent procedure. Furthermore, if a problem can be specified for many data sets, the analysis of all the data sets can be automated, with an attendant reduction in the time required.

The limitation of the analyst's influence to the initial specification of the analysis problem can result in a loss in performance accuracy. Since the procedure is necessarily very general, advantage cannot be taken of peculiarities inherent in a particular data set nor of the analyst's skill in interacting with the analysis process. The loss of performance accuracy must be weighed against the gain in speed and objectivity possible with this automated procedure.

This procedure is a step toward an automated multispectral data analysis system and the rapid analysis of the large volumes of remote sensing data made available by satellite-borne sensor systems. It was synthesized for the purpose of obtaining an objective comparison of multispectral sensor systems, classification criteria, and data sets collected under various environmental circumstances [4]. The effectiveness of the procedure is being rigorously evaluated. Experience has already demonstrated, however, that a price is paid (in terms of overall classification accuracy) when the element of interaction between the skilled analyst and the data analysis process is severely reduced or eliminated. More research in this direction is clearly indicated, either to make the analysis process still "more clever" in emulating the analyst or to reintroduce the analyst to the process in such a way as to maintain overall objectivity.

4. REFERENCES

[1] Phillips, T. L., ed. LARSYS User's Manual. Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana, 1973.

[2] Swain, P. H. "Pattern Recognition: A Basis for Remote Sensing Data Analysis," LARS Information Note 111572, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana, 1972.

[3] Cramér, H. Mathematical Methods of Statistics, Princeton University Press, Princeton, 1946.

[4] Hall, F. G., Bauer, M. E. and Malila, W. A. "First Results from the Crop Identification Technology Assessment for Remote Sensing (CITARS)", Ninth International Symposium on Remote Sensing of the Environment, Ann Arbor, Michigan, 1974.

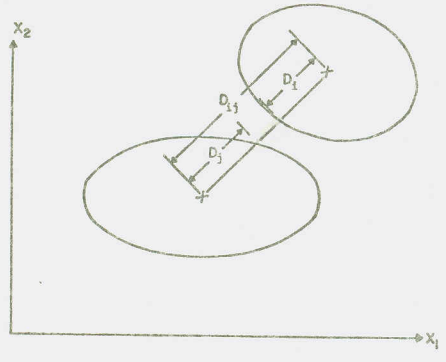

FIGURE 1. SEPARABILITY OF CLUSTERS.