OPTIMIZATION OF A NON-TRADITIONAL UNSUPERVISED CLASSIFICATION APPROACH FOR LAND COVER ANALYSIS

R. K. Boyd, Computer Sciences Corp.
NASA/Goddard Space Flight Center

J. O. Brumfield
Marshall University

W. J. Campbell
NASA/Goddard Space Flight Center

ABSTRACT

The purpose of this paper is to analyze the conditions under which a hybrid of clustering and canonical analysis for image classification produce optimum results. The approach involves generation of classes by clustering for input to canonical analysis. The importance of the number of clusters input and the effect of other parameters of the clustering algorithm (ISOCLS) were examined. The approach derives its final result by clustering the canonically transformed data. Therefore the importance of number of clusters requested in this final stage was also examined. The effect of these variables were studied in terms of the average separability (as measured by transformed divergence) of the final clusters, the transformation matrices resulting from different numbers of input classes, and the accuracy of the final classifications.

The research was performed with Landsat MSS Data over the Hazleton/Berwick Pennsylvania area. Final classifications were compared pixel by pixel with an existing geographic information system to provide an indication of their accuracy.

The results show that both the number of clusters input to canonical analysis and the number of clusters the canonically transformed data is clustered into effect the classification accuracy. Inputting sixty clusters to canonical analysis and clustering the transformed data into thirty clusters provided the best results for the informational categories studied (urban, including commercial/industrial, and residential, agriculture, water, and surface mining) i.e., spectrally very difficult to separate classes.

A definite relationship between the number of clusters input to canonical analysis and the resulting transformation

coefficients was also observed. Specifically, those input numbers of clusters resulting in the highest level of agreement with the GIS Data also produced transformation coefficients most different from those produced by other numbers of input clusters. The separability analysis also tended to support the higher classification accuracies associated with clustering the transformed data into intermediate numbers of clusters as well as the differences associated with the number of clusters input to canonical analysis.

INTRODUCTION

Various authors have reported significant improvements in classification accuracy associated with the use of a non-traditional unsupervised classification procedure. These accuracy improvements have been identified for both areal estimates and pixel by pixel comparisons with ground truth (Brumfield et al., 1981, Witt et al., 1982).

The procedure involves canonical analysis of the statistics derived from an iterative clustering algorithm. The transformation matrix thus developed is used to transform the original data which is then subjected to the same clustering procedures. The procedure provides all of the advantages of using clustering to derive training class statistics (and unsupervised classification in general) (Fleming and Hoffer, 1977) while at the same time incorpoarting the noise reduction and transformation optimization characteristics of canonical analysis (Brumfield et al, 1981).

Although the approach requires very little analyst involvement, decisions must be made regarding the number of classes input to the canonical analysis and the number of classes into which the resulting transformed data should be clustered.

The purpose of this paper is to examine the relationship between these variables and the resulting classification accuracy. Various other indicators of the performance of the procedure are also considered.

DISCUSSION OF METHODS AND PROCEDURES

1. DATA SETS

The remote sensing data used in the experiments were a thirty-four (34) kilometer square subset of the Landsat MSS scene 1350-15190, dated July 8, 1973, covering the Hazleton-Berwick, Pennsylvania area. The MSS data were observed to

exhibit variable haze cover, radiometric striping, and a small amount of random noise. The study site, dissected by the Susquehanna River, is comprised of forested mountains separated by rolling valleys that have been put to a variety of agriculture usages. Major coal mining activities and the associated open pit mines, in all stages of operation, reclamation, and abandonment, are also found in the area. The industrial/commercial activities and residential sprawl of varying densities are also well represented in the area.

Two sets of color infrared photography flown in January and August of 1973 were used as reference data.

A vector (polygon formatted) data base, part of the Environmental and Land Use Data System (ELUDS) of the Pennsylvania Power and Light Company was used as ground truth for performing accuracy assessments. The following categories are coded in the vegetation/landcover layer: urban land, barren land, agricultural land, tree plantations, needle leaf forest, broad leaf forest, mixed forest, scrub land, meadow, forested wetland, unforested wetland, and waterbody.

2. EQUIPMENT

The experiments were carried out using the Interactive Digital Image Manipulation System (IDIMS) (Electromagnetic Systems Laboratory 1981) at the Eastern Regional Remote Sensing Applications Center (ERRSAC), NASA/Goddard Space Flight Center, Greenbelt, MD. This system consists of several components including a Hewlett-Packard Model 3000 mini-computer, a Comtal and Deanza image display terminal, a Talos coordinate digitizer table, and the associated software. The Environmental Systems Research Institute (ESRI) polygon to grid conversion software also played an important role in the research (ESRI, 1979). Canonical analysis was performed by the program CANAL developed by the Office of Remote Sensing For Earth Resources (ORSER) at the Pennsylvania State University (Turner et al, 1978).

3. PREPARATION OF DATA SETS

In order to allow comparison of the MSS data with the landcover information coded in the ELUDS data base the two were altered so as to correspond to a common grid system. Prior to altering the geometric characteristics of the Landsat data a histrogram matching algorithm was applied to remove the six line striping in the data. The Landsat data were then resampled to a grid system referenced to the universal transverse mercator (UTM) map projection (the same map projection ELUDS polygons are referenced to). The transformation coefficients driving the resampling were derived from a third order fit of 30 ground control points

(ordered pairs of Landsat pixel addresses and UTM grid system coordinates). RMS error for these ground control points was less than 0.5 pixel. The cell size of the grid system was chosen to be 67 meters. A gridded version of the ELUDS data base, with the same UTM origin and grid cell size as the Landsat Data was created by determining for each grid cell the data value of the polygon occupying the largest part of the grid cell.

4. INITIAL CLUSTERING

The first step in the procedure is to separate the remote sensing data into spectral clusters for input to the canonical analysis program.

The IDIMS program ISOCLS was used for this step. ISOCLS is a clustering algorithm which either splits or combines clusters in each iteration depending on the requirements set by the analyst for the maximum standard deviation within a cluster (STDMAX) and the minimum euclidean distance between clusters (DLMIN). ISOCLS can be seeded with class means provided by the analyst or with a single cluster defined by the mean vector of the data set to be clustered. In the latter case, this initial cluster is successively split in consecutive iterations until the resulting clusters are less variable than STDMAX. If STDMAX is set low enough the splitting will continue until the maximum number of clusters (also set by the analyst) is met; at which point ISOCLS will iterate assignment of pixels to the clusters and recalculation of the cluster mean vectors until the maximum number of iterations (set by the analyst) is reached. In this way, ISOCLS can be forced to approximate a K-means clustering algorithm (Moik, 1980).

ISOCLS was applied to the entire data set (512 lines by 512 samples). The maximum number of clusters was set to be 10, 20, 30, 40, and 60 in five separate runs. STDMAX was set at 1.5 thus forcing ISOCLS to split the initial clusters until the maximum number of clusters was reached in each case and iterate on that number of clusters as discussed above.

ISOCLS was also applied to supervised (pure) samples of water, strip mines, forest, agriculture, and urban. The supervised samples contained multiple training sites and were selected on the basis of analyst judgement to be as representative of the cover types mentioned as possible. Each sample was clustered separately, and the maximum number of clusters was set at six for each sample, resulting in 30 clusters total. This method of generating classes for input to canonical analysis is not part of the nontraditional unsupervised classification procedure and was included primarily to serve as a point of comparison.

5. DATA TRANSFORMATION

This part of the procedure utilizes a linear transformation of the data. The coefficients for the transformation are calculated by the canonical analysis algorithm developed at ORSER. The algorithm determines the translation, rotation, and rescaling of the data that maximizes the among cluster variability while setting the within cluster variability equal to unity (Merembeck et al., 1978). The resulting canonical transformation maximizes the separability of the clusters based upon the within cluster and among cluster varibility.

The means and convariance matrices for each set of clusters derived from the procedures outlined above were input to the ORSER program CANAL to develop a transformation matrix for each set (Table II). Each transformation matrix was then input to the IDIMS program KLTRANS to perform matrix multiplications with the original data set to generate the transformed data for each case (Brumfield et al., 1981).

6. CLUSTERING OF TRANSFORMED DATA

The final step of the procedure is to classify the transformed data by separating the data into groups with clustering.

The transformed data sets derived from the above procedures were clustered using only the first and second transformed axes (axes one and two contain over 98 percent of the variability in the data). The STDMAX parameter in ISOCLS was set at 0.1, again forcing ISOCLS to emulate a K-means clustering algorithm. ISOCLS was used to generate 15, 20, and 30 clusters for each transformed data set discussed above. ISOCLS was also used to generate 40 clusters for the transformed data set based on 60 clusters. Table I shows the various combinations of clusters input to canonical analysis and output from clustering the transformed data sets. The clusters in each clustered transformed data set were then grouped into informational categories by comparing the cluster results with color infrared photography. Each cluster output was displayed and colored up on a color display screen to effect the comparison. The grouping process was also assisted by examination of two dimensional plots of the cluster means and covariances.

7. SEPARABILITY ANALYSIS

The first indicator used to check for differences related to the number of classes input to and output from the procedure was interclass separability. A modified version of the IDIMS function diverge was used to calculate the average transformed divergence (Swain and Davis, 1978) of those class pairs which yielded transformed divergence values less than 1500 (transformed divergence takes on values between 0 and 2000, where 2000 indicates maximum separability). This average

separability of the least separable classes was calculated for each set of clusters input to the nontraditional unsupervised procedure as well as for each set output from the procedure and is graphed in Figure I.

8. DETERMINATION OF LEVELS OF AGREEMENT

The second indicator of differences associated with the number of classes input to and output from the procedure was level of agreement with ground truth. The land cover layer of the ELUDS Data Base served as ground truth for this study.

The classes in each clustered transformed data set were grouped into five informational categories (urban, strip mines, agriculture, forest, and water) for comparison with the ELUDS landcover information. The grouping was accomplished by renumbering each cluster in each clustered transformed data set to the number chosen to represent the assigned category. The 12 ELUDS landcover classes were grouped into the same informational categories and renumbered to reflect the same coding scheme. Each renumbered clustered transformed data set was then compared pixel by grid cell with the renumbered ELUDS landcover layer to produce a contingency table showing the number of pixels in agreement and disagreement by category. Percentages of agreement were calculated by category and are shown in Table III. Percentage of agreement was calculated by dividing the number of pixels in agreement for the category in question by the total number occurring in the data base for that category. Overall agreement was calculated by dividing the total number of pixels correctly classified by the total number in the data base. These figures are being referred to as levels of agreement instead of accuracy because of the fact that ground verified test sites were not used to calculate them. The ELUDS Data Base is undoubtedly fairly accurate. However, to the knowledge of the authors, no quantitative estimate of its accuracy exists.

RESULTS

1. TRANSFORMATION COEFFICIENTS

The transformation coefficients for Axis 1 and Axis 2 resulting from canonical analysis of the various numbers of input classes are shown in Table II. The coefficients seem to fall into four unique sets, those based on 10 clusters, those based on 20, 30 and 40 clusters, those based on 60 clusters, and those based on the 30 clusters from supervised samples. Without question both the number and source of the class statistics input to canonical analysis affect the resulting transformation coefficients.

2. SEPARABILITY ANALYSIS

The average separability of the least separable classes for each set of input and output clusters is shown in Figure I. Three main trends can be seen from this graph. First, the average separability of the least separable classes tends to increase as the number of clusters increases. Second, for any given number of output classes the average separability of the output classes is constant or decreases slightly as the number of input classes increases. Third, there is a slight increase in separability as the number of output classes is increased for any given number of input classes. Unfortunately, the magnitude of the third trend cannot be viewed as being significant due to the inherent variability associated with calculating transformed divergence from class statistics (Swain and King, 1973). Interestingly, this increased separability of the 60 cluster set of input classes over the 20, 30 and 40 cluster sets (first trend) does concur with the changes observed in the transformation coefficients resulting from those sets.

3. PERCENTAGE OF AGREEMENT

The percentage of agreement of each set of output classes with the ELUDS Data Base is shown in Table III. As is evidenced by the low percentages of agreement for urban and barren, separating these categories from the other categories with MSS Data in this area is very difficult. However, of greater relevance to the scope of this paper are the trends observed in the levels of agreement. Perhaps the most obvious difference is the difference in overall agreement between 15 classes output and 20 or 30 classes output. This decreased overall agreement is consistent with the decreased average separability discussed earlier. The three highest overall levels of agreement were obtained from the 60/30, 60/40, and 30(supervised)/30 sets. Furthermore, with the exception of the 10/20 set, the highest agreement for barren were also obtained with the 60/30, 60/40, and 30(supervised)/30 sets. The results also show that there is an interplay between the number of classes input and the number of classes output. Although, 10 classes input produced an overall level of agreement of 74.7 percent for 20 classes output, it produced only 72.7 percent for 15 classes and 30 classes output. Similarly 60 classes input produced 16.4 and 12.5 percent agreement for barren for 30 and 40 classes output but 0 percent for 20 classes output. Finally, the source of the classes has a definite influence on level of agreement. Thirty input classes from supervised samples produced higher overall agreement than 30 input classes from a systematic sample of the data, regardless of the number of output clusters used in the latter case.

CONCLUSIONS

Clearly the number and source of classes input to and output from the nontraditional unsupervised technique has an impact on the resulting classification accuracy. The results indicate the best overall classification will be obtained when the classes input to canonical analysis sufficiently subdivide the total spectral variability in the data set. In this experiment it was necessary to cluster a systematic sample of the data into 60 clusters or separately cluster supervised samples of the data into six clusters each to accomplish that subdivision. Although certain lower numbers of input classes may produce good results when used in combination with certain other numbers of output classes (e.g. 10/20 in this experiment) it will be difficult to predict these combinations in advance. By subdividing the data set into a large number of clusters the likelihood of representing spectral groupings associated with informational categories is increased. The results also show that separating the transformed data into an intermediate number of clusters is sufficient to obtain the best classification. In this experiment no significant increase in level of agreement was obtained as the number of output classes was increased from 30 to 40. Furthermore, comparable results were obtained when 30 classes were output from the transformed data based on 60 clusters from a systematic sample and from the transformed data based on 30 clusters from supervised samples.

The optimum numbers of clusters will undoubtedly vary from data set to data set. However, it is doubtful that any data set will contain categories more difficult to separate than urban, strip mines, agriculture, and water as contained in the data set used in this experiment. On this basis the 60/30 combination should provide nearly optimal results for any MSS data set.

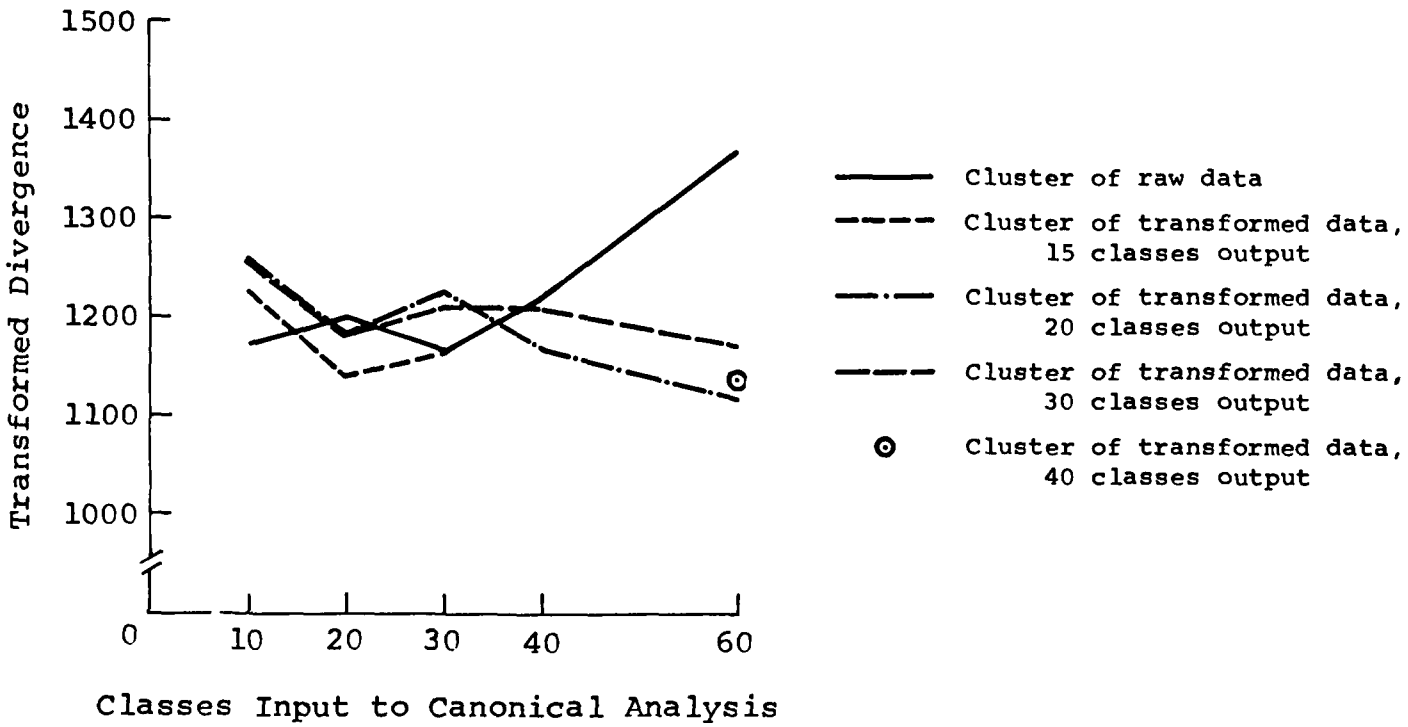

FIGURE I. Average separability of the least separable classes for each set of clusters input to and output from the nontraditional unsupervised classification approach.

TABLE I.

Combinations of number of clusters input to/ and output from the nontraditional classification procedure.

		Input Clusters				
		10	20	30	40	60
Output Clusters	15	X	X	X		
	20	X	X	X	X	X
	30	X	X	X,S	X	X
	40					X

X-Input clusters generated by clustering entire data set.
S-Input clusters generated by clustering supervised samples of the data.

TABLE II.

Transformation Coefficients
for axes one and two, as
produced by canonical analysis.

		Axis 1			
		Band 1	Band 2	Band 3	Band 4
Input Clusters	10	-.1193	-.0260	.0774	.0670
	20	-.2471	-.1096	.1189	.2525
	30	-.2496	-.1686	.0743	.3607
	40	-.2405	-.1639	.0644	.4217
	60	-.1183	-.1610	.2535	.4760
	30[1]	-.0988	-.0344	.1775	.2820
		Axis 2			
		Band 1	Band 2	Band 3	Band 4
	10	-.0320	.1386	.0769	-.0128
	20	.0265	.2879	.1752	-.0142
	30	.0540	.3468	.1794	.0097
	40	.0788	.3538	.2319	-.0182
	60	.2451	.4197	.2394	-.0086
	30[1]	.2096	.2146	.1148	-.0251

[1]From supervised samples.

TABLE III.

Percentage of agreement of classifications with the ELUDS data base.

Input/ Output	Urban	Strip	Agri.	Forest	Water	Overall
10/15	19.8	0.0	57.2	93.3	73.1	72.7
20/15	20.8	0.0	57.4	92.2	72.6	72.1
30/15	22.5	0.0	57.4	93.3	72.7	72.9
10/20	22.3	20.9	75.5	88.2	53.9	74.7
20/20	16.3	0.0	77.5	89.3	70.3	74.5
30/20	16.0	0.0	75.9	90.5	70.5	74.9
40/20	16.0	0.0	77.2	89.4	70.2	74.5
60/20	13.5	0.0	79.1	89.4	67.4	74.7
10/30	18.7	0.8	68.4	89.4	66.8	72.7
20/30	31.3	3.4	60.6	93.5	65.4	74.3
30/30	31.2	3.1	64.1	92.0	66.4	74.2
40/30	30.1	3.5	65.8	91.0	65.1	73.9
60/30	27.1	16.4	72.3	90.1	63.5	75.5
60/40	24.2	12.5	70.7	91.5	65.8	76.1
30[1]/30	25.4	15.8	73.4	90.9	63.8	75.9

[1]Generated by clustering of supervised samples.

REFERENCES

1. Brumfield, J. O., H. L. Bloemer, W. J. Campbell, "An Unsupervised Approach for Analysis of Landsat Data to Monitor Land Reclamation in Belmont County, Ohio", Seventh International Symposium for Machine Processing of Remotely Sensed Data.

2. Electromagnetic Systems Laboratory, "IDIMS Functional Guide", 1981.

3. Environmental Systems Research Institute, "Environmental and Land Use Data System (ELUDS)," 1978.

4. Environmental Systems Research Institute, "Grips User Manual", 1979.

5. Fleming, M. D. and R. M. Hoffer, "Computer Aided Analysis Techniques for an Operational System to Map Forest Lands Utilizing Landsat MSS Data." Laboratory For Applications of Remote Sensing, 1977.

6. Merembeck, F., F. Y. Borden, M. H. Podwysocki, and D. N. Applegate, "Application of Canonical Analysis to Multispectral Scanner Data." 14th Annual Symposium on the Application of Computers in Mine and Industry, 1977.

7. Moik, J. G., "Digital Processing of Remotely Sensed Images." National Aeronautics and Space Administration, 1980.

8. Turner, B. et al., "Satellite and Aircraft Multispectral Scanner Digital Data User Manual," Office for Remote Sensing of Earth Resources, 1978.

9. Swain, P. H., and Davis, S. M., "Remote Sensing The Quantitative Approach." McGraw-Hill International Book Company, 1978.

10. Swain, P. H., and King, R. C., "Two Effective Feature Selection Criteria For Multispectral Remote Sensing." Laboratory for Applications of Remote Sensing, 1973.

11. Witt, R. G., H.H.L. Bloemer, Beldeon Bly, J. O. Brumfield, W. J. Campbell, "Comparing Digital Data Processing Techniques for Surface Mine & Reclamation Monitoring", American Society of Photogrammetry annual meeting, April, 1982.