NASA CR-
ERIM 109600-39-R

Technical Memorandum

THE USE OF UNSUPERVISED CLUSTERING AS A CLASSIFIER FOR LACIE MSS DATA

ALEX P. PENTLAND

OCTOBER 1975

Prepared for
NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

NASA/Johnson Space Center
Contract NAS9-14123

ENVIRONMENTAL
RESEARCH INSTITUTE OF MICHIGAN
FORMERLY WILLOW RUN LABORATORIES, THE UNIVERSITY OF MICHIGAN
BOX 618 • ANN ARBOR • MICHIGAN 48107

ERIM
FORMERLY WILLOW RUN LABORATORIES, THE UNIVERSITY OF MICHIGAN

THE USE OF UNSUPERVISED CLUSTERING AS A CLASSIFIER FOR LACIE MSS DATA

Alex P. Pentland

October 1975

Technical Memorandum
109600-39-R

Prepared Under Contract NAS9-14123
With
NASA/Johnson Space Center

ERIM
FORMERLY WILLOW RUN LABORATORIES, THE UNIVERSITY OF MICHIGAN

Technical Memorandum No. 109600-39-R
20 October 1975

THE USE OF UNSUPERVISED CLUSTERING AS A CLASSIFIER FOR LACIE MSS DATA

Alex P. Pentland

INTRODUCTION

Up to now fairly conventional approaches have been considered for the classification of LACIE data. Of course, more unconventional approaches may be better suited to LACIE and its specific constraints.

During recent months, ERIM has been investigating the use of a clustering algorithm for classification of LANDSAT MSS data, particularly as such classification might apply to the LACIE. This report documents the preliminary results of this continuing investigation.

DESCRIPTION OF THE CLASSIFICATION METHOD

During the performance of NASA contract NAS9-14123 task IV, the ERIM clustering algorithm* was developed to help in the formation of signatures for various classifiers. Because this algorithm is a one-pass algorithm which classifies data points to each cluster, it became obvious that this clustering algorithm could be adapted to perform classification. During the performance of the current contract, differences between conventional LACIE ground truth inputs and the expanded ground truth requirements of the ERIM mixtures algorithms made it necessary to use this clustering algorithm as a classifier in order to obtain suitable 'ground truth' information [1].

This clustering algorithm forms estimates of ground class distributions by classifying each data point as a member of a ground class for which there has already been formed an estimate of distribution, or as a member of a

*See Appendix A for description of the clustering algorithm.

ground class for which no such estimate exists, or the data point is stored, to be classified as one of the above after more information has been gained. In the first case, the estimate of the distribution is modified to reflect the inclusion of the data point, in the second case an estimate of the new class distribution is begun.

When using this clustering algorithm as a classifier the third case, where the data point is stored, cannot be allowed due to storage and data manipulation problems encountered while trying to maintain the data point -- location relationship. This means that we are forced to always make a decision about each data point as it arrives. But when few points have been classified, the chances of an erroneous decision are great, because not enough data points have been received from each distribution to form an accurate estimate of that distribution.

For this reason it is desirable that the mean and variance estimates of each distribution be well established before actual classification begins.

This we may accomplish by "initializing" the clusters -- i.e., we allow the program to cluster an inhomogeneous portion of the scene before starting the classification. In this manner we may be reasonably sure that all major ground classes have contributed enough data points to make an accurate estimate of their distributions.

Because the clusters have no identity (crop type label) associated with them, we must have in the region to be classified an area where ground truth is known. Then, after classification, we may use a classification map of the ground truth area and a transparent overlay of the ground truth to determine the crop type associated with each cluster. This information may be provided by the AI's or by the use of MASC-like signature extension algorithms.

It is in this identification step that the most serious problem arises, for it may be that there are no data points assigned to a particular cluster in the ground truth area, and so we cannot assign an identity to that cluster.

There are two methods for attacking this problem when it arises. First is to use the pairwise probability of misclassification to measure the 'overlap' between clusters in order to associate our problem cluster with clusters of known identity. The drawback to this approach is that the problem cluster may not always overlap with only one crop type. The second method is to examine regions where the problem cluster has data points assigned to it. We may then look for spatial patterns of its occurrence -- for instance, it may occur most in areas that are all of one crop type, or at the edge of areas of one crop type. Because the identity of surrounding data points is now known, these patterns may be discerned with relative ease. The problem with this approach is that such patterns may not always exist, or may be confusing. It has been our experience, however, that these two methods are always sufficient to identify the crop type of each cluster.

The steps involved in carrying out the procedures are given below.

STEP 1 - Cluster over inhomogeneous area thought to contain all major ground classes to initialize estimates of ground classes.

STEP 2 - Use clustering algorithm to classify data set, outputting tape showing cluster classification.

STEP 3 - Use output tape of STEP 2 to produce classification map of ground truth area.

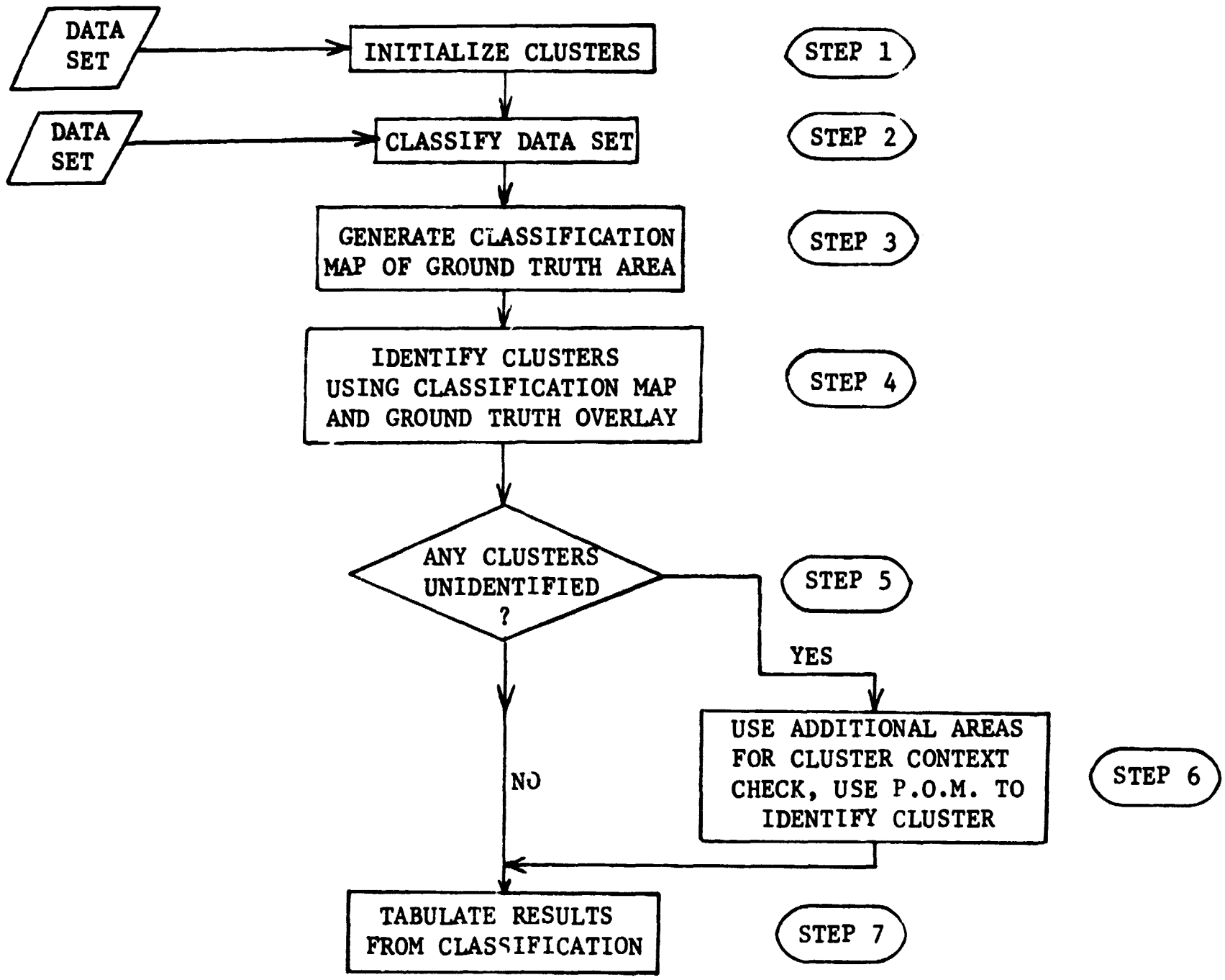
STEP 4 - Using transparent ground truth overlay, identify clusters by means of the frequency with which they appear in various crop type fields.

STEP 5 - If no clusters are left unidentified (or only those with very small populations) proceed to STEP 7.

STEP 6 - Identify areas in which data points assigned to problem clusters appear. Identify the context in which they appear. Find the probability of misclassification of the problem cluster with clusters of known identity. Use these two factors to identify problem cluster.

STEP 7 - Tabulate results.

A flowchart for these steps can be seen in the accompanying figure.



FLOWCHART FOR UNSUPERVISED CLASSIFICATION

PRELIMINARY TEST RESULTS

The first data set chosen for preliminary testing of this method of classification was the Ellis County, Kansas, 12 June data set, one of the LACIE intensive test sites. The method described above was employed with the crop types being 'wheat' and 'other'. The training area was the northern most three sections of this three-by-three section area, with the test area being the remainder. The initialization area was approximately 500 points chosen at random from the test area. In order to determine the crop type of each cluster it was necessary to use both probability of misclassification measures and the spatial context of the clusters, although no severe problems in identification were encountered. Results for test and training areas are presented below.

FIELD CENTER RESULTS

TRAINING				TEST			
Actual Class	Classified as %	Wheat	Other	Actual Class	Classified as %	Wheat	Other
Wheat		93.7	6.3			99.25	0.75
Other		2.55	97.45			3.65	96.35

PROPORTION ESTIMATION RESULTS (Over Whole Area)

	ESTIMATED % WHEAT	GROUND TRUTH % WHEAT
Training	41.5	43.6
Test	45.6	44.5

A second preliminary test was undertaken on the LACIE data set from Randall, Texas, 27 May. To simulate LACIE ground truth provisions, nine training fields were selected at random from this three section by three section data set. These fields included four wheat fields, two corn fields, two summer fallow fields, and one grass field. The class types were again 'wheat' and 'other'. The clustering classification method was then employed, as described above. The initialization area employed in Step 1 was an area of approximately 1000 points chosen at random from an area north of the test site. Because of the fairly scanty ground truth areas, the crop type identification step of the algorithm (Step 6) was more difficult than before.

One major problem was encountered. Field Number 57, which is 303 acres in size, was identified as wheat in the ground truth supplied to us. However, examination of signals from this field showed that they were identical to signals from adjoining 'other' fields. From this it was concluded that Field 57 was, in fact, not wheat, and in computing results it was classed as 'other'.

Results for the test area are presented below.

FIELD CENTER TEST RESULTS

Actual Class	Classified as % Wheat	Other
Wheat	98.09	1.9
Other	1.10	98.89

PROPORTION ESTIMATION RESULTS (Over Whole Area)

	Estimated % Wheat	Ground Truth % Wheat
Test	30.61	30.01

On the basis of these two preliminary tests, it was concluded that this method of classification was accurate enough to be useful in 'extending' ground truth, and so was included in the current Test and Evaluation task for that purpose.

FURTHER TESTS

Four LACIE intensive test sites were then processed with this method, including the two sites reported above which were reprocessed, using simulated LACIE ground truth.* In each case the initialization area consisted of approximately 1000 points from north of the test sites, and the crop types were 'wheat' and 'other'. Virtually no cluster identification problems were encountered in any of the test sites, although probability of misclassification measures and spatial context were used in each site.

From each site several sections were selected at random, and the proportion of wheat was estimated for these sections. Although the field center results were not tabulated, in each case the diagonal terms of the performance matrix appeared to be well above 90%. The results of proportion estimation on these four sites are presented below.

Intensive Test Site	Estimated Wheat	Actual Wheat	RMS Error Sec. by Sec.	Number of Sections Estimated
Deaf Smith, Texas, 27 May	.331	.333	.06	4
Ellis, Kansas, 12 June	.504	.458	.046	4
Randall, Texas, 27 May	.455	.472	.033	5
Finney, Kansas, 26 May	.222	.2066	.027	9

*See ERIM T&E Quarterly Report [3] for description of the selection of ground truth fields.

CONCLUSIONS

Examination of these results shows that this method of classification appears to give accurate field center results and, more importantly, to give practical, statistically consistent and accurate estimate of crop proportions.

The accuracy of this method appears to be attributable to certain qualities of the particular clustering algorithm used. These qualities are freedom from assumptions about Gaussian data, and the continual updating of distribution estimates, including updating the number of modes.

We have also found that this method is relatively tolerant of errors in the determination of crop type, as crop identity is used only for identifying clusters, and not for computing signatures.

RECOMMENDATIONS

We feel that these test results show that this method of classification deserves additional investigation and development for possible inclusion into the LACIE project.

ACKNOWLEDGEMENTS

The author would like to acknowledge the administrative direction provided by Richard F. Nalepka.

APPENDIX A
DESCRIPTION OF THE CLUSTERING ALGORITHM

This algorithm [2] uses small, normal distributions as elements with which to approximate the cumulative distribution function of the ground classes in a scene. A description of this follows.

(1) Suppose we have M cells $\Gamma_1 \ldots \Gamma_m$, each with mean M_i, variances $(\sigma_{i\,1}^2 \ldots \sigma_{i\,N}^2)$, where N = number of channels and K_i = number of samples within the cell. Given a new sample X, calculate the distance of X from each cell center

$$d(X,M_i) = \sum_{j=i}^{N} (X_i - M_{ij})/\sigma_{ij}^2 \quad (i=1,\ldots M)$$

Find K such that

$$d(X,M_K) = \text{MIN}_i\ d(X,M_i)$$

Then X is classified as one of the following:

$$d(X,M_K) \leq \tau \rightarrow X \text{ assigned to } \Gamma_K$$

$$d(X,M_K) \geq \theta \rightarrow X \text{ creates a new cell,}$$

otherwise X is stored.

(2) When a new sample is classified to the i^{th} cell, this cell's parameters are adjusted as follows:

(a) Increase number of samples (K_i) by one

(b) Calculate new mean vector (M_i)

$$M_i = \frac{1}{K_i} \sum_{\ell=1}^{K_i} X_{i,\ell}$$

(c) Determine new variances by

$$\sigma^2_{i,j} = \text{MAX}(\sigma^2_{i,j}(0), S^2_{i,j})$$

where

$$S^2_{i,j} = \frac{1}{K_i} \sum_{\ell=1}^{K_i} (K_{\ell j} - M_{ij})^2$$

where the X_ℓ are classified to the i^{th} cell and $\sigma^2_{ij}(0)$ is an initial assignment of σ^2_{ij}. Only when S^2_{ij} exceeds $\sigma^2_{ij}(0)$ do we replace $\sigma^2_{ij}(0)$ with S^2_{ij}.

(3) The first sample always creates a new cell. The second sample is tested and classified by (1) and so on.

When all samples have been classified, the stored samples are forced into the nearest cells according to (1).

REFERENCES

[1] Estimating Proportions of Objects From Multispectral Scanner Data, H. M. Horwitz, J. T. Lewis, A. P. Pentland, NASA Contract NAS9-14123, Task IV, ERIM Report 109600-13-F, May 1975.

[2] "An Algorithm for Nonparametric Pattern Recognition", G. E. Sebestyen, and J. Edie, IEEE Trans. Electronic Computers EL-15, 908-916 (Chapter 6) 1966.

[3] Evaluation of Algorithms for Wheat Inventory, H. Horwitz, R. Crane, W. Richardson, J. Lewis, J. Reyer, A. Pentland, NASA Contract NAS9-14123, Task 14, ERIM Report 109600-37-L, August 1975.