CLASSY—An Adaptive Maximum Likelihood Clustering Algorithm*

R. K. Lennington[a] *and M. E. Rassbach*[b]

ABSTRACT

A new clustering method called CLASSY, which alternates maximum likelihood iterative techniques for estimating the parameters of a mixture distribution with an adaptive procedure for splitting, combining, and eliminating the resultant components of the mixture, has been developed. The adaptive procedure is based on maximizing the fit of a mixture of multivariate normal distributions to the observed data using its first through fourth central moments. The method generates estimates of the number of multivariate normal components in the mixture and the proportion, mean vector, and covariance matrix for each component.

This paper describes the mathematical model which is the basis for CLASSY and outlines the actual operation of the algorithm as currently implemented. Results of applying CLASSY to real and simulated Landsat data are presented and compared with results generated by the Iterative Self-Organizing Clustering System (ISOCLS) algorithm, a derivative of the ISODATA algorithm, on the same data sets.

INTRODUCTION

The Large Area Crop Inventory Experiment (LACIE) is dependent on clustering for the determination of spectral classes within a Landsat image of a sample segment (ref. 1). Currently, the Iterative Self-Organizing Clustering System (ISOCLS) is used for this purpose (refs. 2 and 3). ISOCLS is basically a variation of the k-means or ISODATA algorithm of Ball and Hall (refs. 4 and 5). Although this algorithm may be interpreted as a simplified maximum likelihood procedure, it is fundamentally a heuristic algorithm for breaking a data set into fairly homogeneous compact clusters.

A new clustering algorithm called CLASSY, which approximates the mixture distribution of a given data set such as Landsat data with a linear combination of normal distributions, has been developed. CLASSY operates by interleaving maximum likelihood iterative estimation with an adaptive procedure for splitting, combining, and eliminating the resultant components of the mixture density (or clusters). The adaptive procedure is based on maximizing the fit of a mixture of multivariate normal distributions to the observed data using its first through fourth central moments. This procedure allows new components (or clusters) to be created if any existing one appears to be multimodal or otherwise nonnormal. CLASSY produces an estimate of the proportion, mean vector, and covariance matrix for each component in the multivariate normal mixture. It differs from standard maximum likelihood procedures in that it also generates an estimate of the number of components in the mixture.

The CLASSY algorithm is currently implemented on an IBM 370-148 computer. It is written in Fortran IV language and currently accepts as input Landsat imagery on magnetic tape. Both line printer and magnetic tape output are generated by the program.

The following section of this paper describes the mathematical model that is the basis for CLASSY and provides a brief description of the actual operation of the algorithm. The section entitled "Results" contains comparisons of the performances of CLASSY and ISOCLS on simulated data and on actual Landsat data used in LACIE. Finally, these results are evaluated and conclusions are developed.

*The current material for this paper was developed under NASA contract NAS 9-15200 and prepared for the Earth Observations Division, NASA Johnson Space Center, Houston, Texas. CLASSY was developed by M. E. Rassbach while he was a National Research Council postdoctoral fellow working at the Johnson Space Center.

[a] Lockheed Electronics Company, Houston, Texas.

[b] Elogic, Inc., Houston, Texas.

MATHEMATICAL DESCRIPTION

Assumptions and Problem Definition

The fundamental mathematical assumption underlying CLASSY is that the data may be usefully approximated by a mixture of multivariate normal densities. That is, if **x** is an observation vector and p is its probability density function, then

$$p\left(\mathbf{x} \,|\, m,\pi_m\right) = \sum_{i=1}^{m} a p_i\left(\mathbf{x} \,|\, \mu_i,\Sigma_i\right) \tag{1}$$

where a_i is the a priori probability of occurrence of class i; $p_i(\mathbf{x}\,|\,\boldsymbol{\mu}_i,\Sigma_i)$ is the multivariate normal probability density function for class i with mean vector $\boldsymbol{\mu}_i$ and covariance matrix Σ_i; m is the total number of classes; $\boldsymbol{\pi}_m$ is the full set of parameters (i.e., $\{a_1, \ldots, a_m, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m, \Sigma_1, \ldots, \Sigma_m\}$).

Given a set of statistically independent, unlabeled sample vectors $\{\mathbf{x}_j\}$, the likelihood function may be formed in the following manner:

$$L\left(\{\mathbf{x}_j\} \,|\, m,\pi_m\right) = \prod_{j=1}^{N}\left[\sum_{i=1}^{m} a p_i\left(\mathbf{x}_j \,|\, \mu_i,\Sigma_i\right)\right] \tag{2}$$

where N is the total number of samples.

So far, the assumptions and equations parallel the usual maximum likelihood development. CLASSY makes the additional assumption that each value of the parameters m and $\boldsymbol{\pi}_m$ occurs with an a priori probability distribution $A\,(m,\boldsymbol{\pi}_m)$. This Bayesian formulation of the problem is taken to avoid the degenerate situation of increasing the likelihood by generating more and more clusters with smaller and smaller values of a_i. The practical limit of this process is that each class will be associated with only one data point.

The objective of CLASSY, then, is to determine the discrete parameter m and the continuous parameter vector $\boldsymbol{\pi}_m$ so as to maximize the following function:

$$L\left(\{\mathbf{x}_j\}\, m,\pi_m\right) = A\left(m,\pi_m\right) \prod_{j=1}^{N}\left[\sum_{i=1}^{m} a p_i\left(\mathbf{x}_j \,|\, \mu_i,\Sigma_i\right)\right] \tag{3}$$

The values of m and $\boldsymbol{\pi}_m$ which maximize equation (3) specify a set of distributions that will be called clusters. Of course, $A(m,\boldsymbol{\pi}_m)$ must be chosen so that it satisfies the normalization constraint

$$\sum_{m=1}^{\infty} \int A\left(m,\pi_m\right) d\pi_m = 1 \tag{4}$$

The upper limit on m is infinity since the possibility of generating an infinite number of clusters must be considered (in theory).

Typically, in the absence of other information, the a priori probabilities may be chosen as

$$A\left(m,\pi_m\right) = \begin{cases} \beta \prod_{i=1}^{m} C_i, \ \pi_m \in R_m \\ 0, \text{ otherwise} \end{cases} \tag{5}$$

where $C_i = C$ is a constant containing normalization factors over $\boldsymbol{\pi}_m$ space, β is an overall normalization constant, and R_m is a finite region of $\boldsymbol{\pi}_m$ space corresponding to allowable values for the parameters. Using this simple form for $A(m,\boldsymbol{\pi}_m)$ in equation (4), the following is obtained.

$$\sum_{m=1}^{\infty} \int_{R_m} \beta C^m \, d\pi_m = \beta \sum_{m=1}^{\infty} \left(C \int_{R_1} d\pi_1\right)^m \tag{6}$$

Now if

$$C = \gamma \left(\int_{R_1} d\pi_1 \right)^{-1}$$

where $\gamma < 1$, then the sum in equation (6) will converge and $\beta = 1 - \gamma$ provides the proper normalization. Thus, larger values of γ provide a priori bias in favor of more clusters, whereas smaller values provide bias in favor of fewer clusters.

In the current version of CLASSY, the authors have been using $\gamma = e^{-1}$ and approximating the R_1 integral of $d\pi_1$ by e^{2d}. This represents a crude approach to the problem of determining the form of $A(m,\pi_m)$. However, in practice, the overall technique to be described in the next section has proven not to be sensitive to reasonable changes in the value of C.

With the form for $A(m,\pi_m)$ assumed in equation (5), the function to be maximized becomes

$$L(\{x_j\}, m, \pi_m) = \begin{cases} \left(\beta \prod_{i=1}^{m} C_i\right) \prod_{j=1}^{N} \left\{ \sum_{i=1}^{m} \frac{a_i}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} \right. \\ \left. \exp\left[- \frac{(x_j - \mu_i)^T \Sigma_i^{-1} (x_j - \mu_i)}{2} \right] \right\}, & \pi_m \in R_m \\ 0, \text{ otherwise} \end{cases} \tag{7}$$

where d is the dimensionality of the observations $\mathbf{x}_j$.

Solution Procedure

Many approaches may be taken to maximize equation (3). The approach chosen in CLASSY is to interleave maximum likelihood iteration (designed to maximize $L(\{\mathbf{x}_j\},m,\pi_m)$ with respect to the continuous parameter vector π_m) with a discrete split, join, and combine process (designed to maximize $L(\{\mathbf{x}_j\},m,\pi_m)$ with respect to the discrete parameter m). Although the theoretical convergence properties of this procedure have not been examined, it is expected that, by alternating these two techniques, values of m and π_m corresponding to at least a local maximum of $L(\{\mathbf{x}_j\},m,\pi_m)$ will be determined. Because the splitting and combining techniques operate around each existing cluster and the statistics for hypotheses concerning different numbers of clusters are maintained separately, it has been observed that the final local maximum will often be global.

Necessary conditions for a maximum of $L(\{\mathbf{x}_j\},m,\pi_m)$ with respect to π_m, assuming a fixed number of classes m, are well known (see Duda and Hart (ref. 6) and Wolfe (ref. 7)) and are given by the following equations:

$$p(i|x_k,\pi_m) = \frac{a_i p_i(x_k|\mu_i,\Sigma_i)}{\sum_{j=1}^{m} a_j p_j(x_k|\mu_j,\Sigma_j)} \tag{8}$$

$$a_i = \frac{1}{N} \sum_{k=1}^{N} p(i|x_k,\pi_m) \tag{9}$$

$$\mu_i = \frac{\sum_{k=1}^{N} p(i|x_k,\pi_m)\, x_k}{\sum_{k=1}^{N} p(i|x_k,\pi_m)} \tag{10}$$

$$\Sigma_i = \frac{\sum_{k=1}^{N} p(i|x_k,\pi_m)\,(x_k - \mu_i)(x_k - \mu_i)^T}{\sum_{k=1}^{N} p(i|x_k,\pi_m)} \tag{11}$$

where $p(i|\mathbf{x}_k, \pi_m)$ is the posterior probability of class i, given the kth sample vector and the values of the parameters, and a_i, μ_i, and Σ_i, $i = 1, \ldots, m$, are the elements of π_m.

Numerous techniques have been proposed for obtaining a solution to this set of coupled, simultaneous nonlinear equations. Specific methods have been suggested by Quirein and Trichel (ref. 8), Day (ref. 9), Hasselblad (ref. 10), and Wolfe (ref. 7), among others. CLASSY uses direct functional iteration for equations (10) and (11); that is, use of estimates for

μ_i and Σ_i on the right side to produce improved estimates on the left side.

Estimates for the a priori class probabilities a_i are computed using an iteration scheme which has proved to converge more rapidly than simple functional iteration using equation (9). The scheme used is specified by the following equation, which is derived in the appendix.

$$a_i = \frac{a_i \sum\limits_{p_i>q_i} \frac{p_i - q_i}{p}}{N - \sum\limits_{p_i>q_i} \frac{q_i}{p} - \sum\limits_{p_i<q_i} \frac{p_i}{p}} \qquad (12)$$

where

$$p_i = p_i\left(\mathbf{x}_k \middle| \mu_i, \Sigma_i\right)$$

$$p = \sum_{j=1}^{m} a_j p_j\left(\mathbf{x}_k \middle| \mu_j, \Sigma_j\right)$$

$$q_i = \sum_{j \neq 1} \left(\frac{a_j}{1 - a_i}\right) p_j\left(\mathbf{x}_k \middle| \mu_j, \Sigma_j\right)$$

N = the total number of observations

This equation is used by substituting old values of a_i, μ_i, and Σ_i, $i = 1, \ldots, m$, on the right to obtain an updated estimate for a_i on the left. The summations are taken over all values of $\mathbf{x}_k$ such that $p_i > q_i$ or $p_i < q_i$.

Initially, each new data point $\mathbf{x}_j$ is used to update the parameter values using equations (8) through (12). This procedure allows rapid evolvement of the parameters as new data points are processed. A danger lies in the fact that the data are considered sequentially. If significant correlation is present in the data, updating the parameters with each new data point could theoretically cause the maximum likelihood equations to converge very slowly or to undergo cyclic drifts. This problem has been found to be particularly severe in Landsat data, which exhibit high correlation within fields. To reduce the effects of this correlation, the data are initially scrambled in a random fashion. Using scrambled data and updating the parameter values with each new data point, the authors have observed that the number of samples (N) required for initial convergence is on the order of a few hundred, even for large data sets. Following initial convergence, the parameters are updated only after a complete pass has been made through the data. This second type of iteration allows a fine tuning of the parameter values and is not subject to problems related to data correlation. The conditions under which the second mode of parameter iteration is entered are discussed later in this section.

The same iteration scheme used to update the parameters is also used to accumulate third- and fourth-order central moments. That is, current values of the parameters are used with each new data point to form the new terms to be accumulated for estimating the moments. The fundamental equations for the estimates of the third- and fourth-order moments are generalizations of equations (10) and (11) and are given as

$$S^{(i)}_{kpq} = \frac{1}{W_i} \sum_{j=1}^{N} \bar{x}_{jk}\bar{x}_{jp}\bar{x}_{jq}\, p\left(i \middle| \mathbf{x}_j, \pi_m\right) \qquad (13)$$

and

$$K^{(i)}_{klpq} = \frac{1}{W_i} \sum_{j=1}^{N} \bar{x}_{jk}\bar{x}_{jl}\bar{x}_{jp}\bar{x}_{jq}\, p\left(i \middle| \mathbf{x}_j, \pi_m\right) \qquad (14)$$

where $\bar{x}_{jk}$ = $(x_{jk} - \mu_{ik})$

x_{jk} = the kth component of the jth sample vector

μ_{ik} = the current estimate for the kth component of the mean vector of cluster i

and where

$$W_i = \sum_{j=1}^{N} p\left(i \middle| \mathbf{x}_j, \pi_m\right) \qquad (15)$$

The parameter W_i is defined as the weight for cluster i and may be considered as the number of points assigned to a cluster on a fractional probabilistic basis; $S^{(i)}$ is a three-dimensional "skewness" tensor, and $K^{(i)}$ is a four-dimensional "kurtosis" tensor. To reduce the number of parameters to be estimated and stored, traces of these tensors are formed using the inverse of the estimated sample covariance matrix for cluster i (Σ_i) to obtain

$$s_k^{(i)} = \frac{1}{W}\sum_{j=1}^{N} \bar{x}_{jk}\left(\bar{x}_j^T\Sigma_i^{-1}\bar{x}_j\right) p\left(i\middle|x_j,\pi_m\right) \quad (16)$$

where $k = 1, 2, \ldots, d$, and

$$K_{kl}^{(i)} = \frac{1}{W}\sum_{j=1}^{N} \bar{x}_{jk}\bar{x}_{jl}\left(\bar{x}_j^T\Sigma_i^{-1}\bar{x}_j\right) p\left(i\middle|x_j,\pi_m\right) \quad (17)$$

where $k,l = 1, 2, \ldots, d$, and

$$\bar{x}_j^T = \left[\bar{x}_{j1}\cdots\bar{x}_{jd}\right] \quad .$$

During the initial iteration mode, when parameter values are changing with each data point, the estimates for

$$S^{(i)} = \left(s_1^{(i)},\ldots,s_d^{(i)}\right)$$

and

$$K^{(i)} = \left(K_{kl}^{(i)}\right)$$

for each cluster i are only approximately correct. The second mode of iteration produces a more accurate estimate of these statistics. As shall be seen, the estimates of $S^{(i)}$ and $K^{(i)}$ are used in the maximization of the likelihood with respect to the discrete parameter m.

The optimization of $L(\{x_j\},m,\pi_m)$ with respect to the discrete parameter m takes the form of generating hypotheses concerning the number of clusters and the subsequent testing of these hypotheses using a likelihood ratio test. At certain points in the process of maximum likelihood iteration, it is possible to generate a hypothesis concerning the fit of a given cluster to the data; namely, either that the data are better represented by two clusters rather than one (a split hypothesis) or that the data are better represented by combining the given cluster with another cluster (a join hypothesis). Each cluster is checked to determine whether either a split or a join hypothesis seems reasonable when the weight for that cluster as defined in equation (15) exceeds a threshold. At this same time, a portion of the old data, which have been accumulated using less accurate parameter values, is subtracted from the appropriate sum for each of the parameters given in equations (8) through (11). The weight threshold is initially set at 200 and increases each time it is exceeded. This procedure allows an initial fit to the major clusters in the data and a subsequent development of more detailed cluster structure.

The generation of a split hypothesis is governed by comparing scalar measures of multivariate skewness and kurtosis for each cluster to thresholds derived from the appropriate distribution for these measures computed under the assumption of a multivariate normal distribution. The scalar measures of multivariate skewness and kurtosis are contractions of the skewness vector $S^{(i)}$ and the kurtosis matrix $K^{(i)}$ with respect to the inverse of the estimated covariance matrix for cluster i, Σ_i^{-1}. These measures are given by

$$s_i^2 = S^{(i)T}\Sigma_i^{-1}S^{(i)} \quad (18)$$

$$k_i = \mathrm{Tr}\left(K^{(i)}\Sigma_i^{-1}\right) \quad (19)$$

$$\left(k_i^\circ\right)^2 = \mathrm{Tr}\left(K^{(i)}\Sigma_i^{-1}K^{(i)}\Sigma_i^{-1}\right) - \frac{k_i^2}{d} \quad (20)$$

Here, k_i is the trace of the normalized kurtosis matrix for cluster i and $(k_i^\circ)^2$ is the trace-free component of the square of k_i.

If any one of these three statistics given by equations (18) to (20) exceeds its threshold value, the hypothesis is formed that the ith cluster may be split into two parts. The parameters for each of the two new component clusters are estimated by minimizing the squared differences between the observed covariance matrix, the skewness vector, and the kurtosis matrix and the corresponding quantities for the mixture distribution composed of the two new normal distributions. The proportion and mean for the mixture composed of the subclusters are defined to be exactly equal to the corresponding quantities for the parent cluster. That is, if a_i and μ_i are the current estimates of proportion and mean for cluster i and a_{i_1}, a_{i_2}, μ_{i_1}, and μ_{i_2} are the corresponding initial values of the subcluster parameters, it is required that

$$a_i = a_{i_1} + a_{i_2} \tag{21}$$

and

$$\mu_i = \frac{a_{i_1}\mu_{i_1} + a_{i_2}\mu_{i_2}}{\hat{a}_{i_1} + \hat{a}_{i_2}} \tag{22}$$

Thus, the difference in subcluster proportions and the difference in the subcluster mean vectors are left as free parameters. The other free parameters are the independent elements of the two subcluster covariance matrices. Therefore, a total of

$$1 + d + 2\left[\frac{d(d+1)}{2}\right] = (d+1)^2$$

parameters must be determined.

There are $[d(d+1)]/2$ equations, each of which matches the covariance matrix and kurtosis matrix parameters for the parent cluster to the corresponding parameters for the subcluster mixture. In addition, there are d equations matching the skewness vector parameters for the parent cluster and the subcluster mixture. This is a total of $d^2 + 2d$ equations. Thus, there is one more free parameter or unknown than there are equations and a unique solution is not possible.

The approach taken to obtaining a solution is to minimize by means of a steepest descent algorithm a quadratic form that may be expressed as

$$\phi = \alpha_1\left\|\Sigma_i \quad \Sigma_p\right\|^2 + \alpha_2\left\|\mathbf{K}^{(i)} \quad \mathbf{K}_p\right\|^2 + \alpha_3\left\|\mathbf{S}^{(i)} - \mathbf{S}_p\right\|^2 \tag{23}$$

where Σ_i, $\mathbf{K}^{(i)}$, and $\mathbf{S}^{(i)}$ are the current estimates of the covariance matrix, the kurtosis matrix, and the skewness vector, respectively, for cluster i; Σ_p, $\mathbf{K}_p$, and $\mathbf{S}_p$ are the corresponding "pooled" estimates from the mixture of the subclusters under the restrictions of equations (21) and (22); and α_1, α_2, and α_3 are arbitrary constants. The norms are the appropriate matrix and vector norms. That is, if M_i is one of the symmetric matrices in equation (23) and $V_i = \mathbf{S}^{(i)} - \mathbf{S}_p$, then

$$\left\|M_i\right\|^2 = \mathrm{Tr}\left(M_i\Sigma_i^{-1}M_i\Sigma_i^{-1}\right)$$

$$\left\|V_i\right\|^2 = V_i^T\Sigma_i^{-1}V_i$$

Minimization of equation (23) under the restrictions of equations (21) and (22) produces estimates for the proportions, mean vectors, and covariance matrices which define two new multivariate normal clusters. In the generation of a split hypothesis, the statistics defining the multivariate normal parent cluster are not discarded. When the maximum likelihood iteration cycle is begun again, it is performed for the previously existing clusters, including the parent cluster, and for the two new clusters, which may be thought of as subclusters of the parent cluster. Thus, as split and join hypotheses are generated, a hierarchical cluster structure or cluster tree evolves. Final decisions concerning the choice of a parent cluster or its subclusters to represent the data are made on the basis of likelihood ratio tests as will be described later.

The generation of a join hypothesis is the inverse of the split hypothesis generation procedure. That is, if the generation of a join hypothesis for two already existing clusters is deemed reasonable, then statistics for a new parent cluster are calculated from the multivariate normal mixture distribution defined by the two clusters to be joined. The new parent cluster is inserted at the level of the clusters to be joined and the clusters to be joined are moved to the next lower level in the tree as subclusters of the new parent.

It should be noted that only clusters which have a common parent are eligible to be joined. The test for determining when a join hypothesis should be generated is designed to measure the degree of overlap between clusters having a common parent cluster. (All the clusters at the top level of the tree are assumed to have a common parent.) The overlap is checked by comparing the mean vectors and the diagonal elements of the covariance matrices for two clusters. A heuristic criterion is used to perform this check. This criterion is given by equation (24).

$$R_{ij} = \frac{(\mu_i - \mu_j)^T \left(\frac{W_i\Sigma_i^{-1} + W_j\Sigma_j^{-1}}{W_i + W_j}\right)(\mu_i - \mu_j) + A\sum_{k=1}^{d}\left(\ln|\sigma_{kk_i}| - \ln|\sigma_{kk_j}|\right)^2}{B\left(\frac{W_j}{W_i} - \frac{W_i}{W_j}\right)^2 + 1} \tag{24}$$

where W_i is the current weight for cluster i and A and B are arbitrary constants (currently, $A = 0.3$ and $B = 0.18$).

The first term in the numerator is a weighted distance between the mean vectors of clusters i and j. The weighting is accomplished by an average inverse covariance matrix for clusters i and j. The second term in the numerator is a measure of the difference in the diagonal elements of the two covariance matrices. The diagonal elements rather than the full covariance matrices are used for computational simplicity. A more complete expression involving all covariance terms would be ln[det $\Sigma_i\Sigma_j^{-1}$]. The denominator is designed to discriminate against small clusters in the sense that R_{ij} will be artificially reduced if the weight of one cluster is small relative to the weight of the other cluster. This factor is designed to give large clusters an opportunity to absorb small clusters if such a join does not substantially affect the statistics of the larger cluster.

The R_{ij} criterion is computed for each cluster having the same parent as cluster i. If the cluster j for which R_{ij} is a minimum is less than an empirically set fixed threshold, then a join hypothesis for cluster i and j is generated.

Final decisions concerning the acceptance or rejection of split and join hyotheses are made in terms of likelihood ratio tests. If there are m_i subclusters for a given parent cluster i, then the logarithm of the likelihood ratio of the subclusters to the parent is accumulated at the same time that maximum likelihood iteration is taking place. The form of this likelihood ratio is given by equation (25).

$$\begin{aligned} \ln \Lambda_i &= \ln\left\{\frac{\beta C^{m_i}\prod_{j=1}^{N}\left[\sum_{k=1}^{m_i} a_{k_i} p\left(x_j|\mu_{k_i}, \Sigma_{k_i}\right)\right]}{\beta C\sum_{j=1}^{N} a_i p\left(x_j|\mu_i, \Sigma_i\right)}\right\} \\ &= \left(m_i - 1\right)\ln C + \sum_{j=1}^{N}\left\{\ln\left[\sum_{k=1}^{m_i} a_{k_i} p\left(x_j|\mu_{k_i}, \Sigma_{k_i}\right)\right]\right. \\ &\quad \left. - \ln\left[a_i p\left(x_j|\mu_i, \Sigma_i\right)\right]\right\} \end{aligned} \tag{25}$$

where Λ_i is the likelihood ratio for cluster i; a_i, μ_i, and Σ_i are the current estimates of the parameters for cluster i; and a_{k_i} and Σ_{k_i} are the corresponding subcluster parameters. This log likelihood ratio is tested against a threshold computed assuming that 2 ln Λ_i is approximately distributed as an x^2 random variable with degrees of freedom equal to $d + 1$. A one-tailed test is used, and the probability of a type I error is set at 0.01. If 2 ln Λ_i exceeds the threshold set by the test, then the statistics for the parent cluster are eliminated and the subclusters take the place of the parent cluster.

It is also possible that ln Λ_i may become negative, even though in theory this should not occur. In practice, negative values may occur because of poor initial estimation of the subcluster parameters or lack of convergence in these estimates. To avoid the expense of maintaining poor subclusters, the subclusters are eliminated in favor of the parent cluster when ln Λ_i falls below a fixed negative threshold. This threshold is set to a large negative value to allow the subcluster statistics to converge if they are going to converge.

One other possibility in testing the likelihood ratio is that the subcluster statistics may actually converge so that the mixture distribution defined by the subcluster parameters reproduces or very nearly reproduces the parent cluster distribution. In such cases, $\ln \Lambda_i$ will remain at a low value possibly slightly greater than or less than zero. If this occurs, it may be assumed that the parent cluster is the most economical description of the data and the subclusters may be eliminated. To test for this situation, another statistic based on the accumulated point probabilities under the parent and subcluster hypotheses is examined. Defining

$$p_{i_i}\left(x_j\right) = \sum_{k=1}^{m_i} a_{k_i} p\left(x_j \middle| \mu_{k_i}, \Sigma_{k_i}\right)$$

where a_{k_i}, μ_{k_i}, and Σ_{k_i} are the current estimates of the parameters for the subclusters of cluster i, the statistic computed is

$$E_i = \sum_{j=1}^{N} \left\{ \frac{p_i\left(x_j \middle| \mu_i, \Sigma_i\right) - p_{i_i}\left(x_j\right)}{p_i\left(x_j \middle| \mu_i, \Sigma_i\right) + p_{i_i}\left(x_j\right)} \right\}^2 \qquad (26)$$

Equation (26) gives a crude measure of how much a parent cluster differs from the mixture of its subclasses. If E_i becomes smaller than a fixed empirically determined threshold and the log likelihood ratio is less than a fixed small positive value, then the subclusters are eliminated in favor of the parent cluster.

The one remaining test in the portion of the program that performs maximization with respect to the number of classes is a simpie test on the proportion a_i of each cluster or subcluster. If this proportion falls below a threshold value, currently set to 0.01, then the cluster is eliminated. This test is used primarily in the interest of efficiency since very small clusters do not significantly affect the overall mixture distribution.

All the tests for the generation of hypothesized new clusters and for the elimination of clusters or subclusters occur at certain intervals during the process of maximum likelihood iteration and statistics accumulation; namely, when the weight for a given cluster has increased by a fixed amount or when a complete pass has been made through the data since the last tests were performed. After the tests have been made and any resultant restructuring of the cluster tree has taken place, E_i (given by eq. (26)), $\mathbf{K}_i$, $\mathbf{S}_i$, and Λ_i are reset. Thus, these statistics depend only on the data processed since the last testing of the cluster statistics for cluster i.

The present program cycles through the data a fixed number of times. (The number of passes through the data is controlled by an external parameter.) When the desired number of passes is complete, the program clusters the data by examining it point by point and assigning each data point to the cluster in the cluster tree for which the probability of occurrence of this data point is the greatest. This is the only time in the program that points are assigned to clusters. When all the points have been assigned, a cluster map showing the cluster symbol for each point is printed out. The program also prints out the final values for the parameters for each cluster in the cluster tree.

Figure 1 is a general flow diagram for the CLASSY program. This is not a detailed flow diagram for the program but merely serves to summarize the information given in this section in a convenient manner.

The initial values assumed at the beginning of the program are as follows.

$$\left.\begin{aligned} m &= 1 \\ a_1 &= 1 \\ \mu_1 &= \begin{bmatrix} 0.04 \\ \cdot \\ \cdot \\ \cdot \\ 0.04 \end{bmatrix} \\ \Sigma_1 &= \begin{bmatrix} 10 & & & 0 \\ & \cdot & & \\ & & \cdot & \\ & & & \cdot \\ 0 & & & 10 \end{bmatrix} \end{aligned}\right\} \qquad (27)$$

DATA, PROCEDURES, AND RESULTS

To evaluate the CLASSY clustering algorithm, it was applied to both real and simulated Landsat data. Performance measures were defined and calculated for each trial of the algorithm. The measures were compared with those derived from applying the ISOCLS algorithm to the same data.

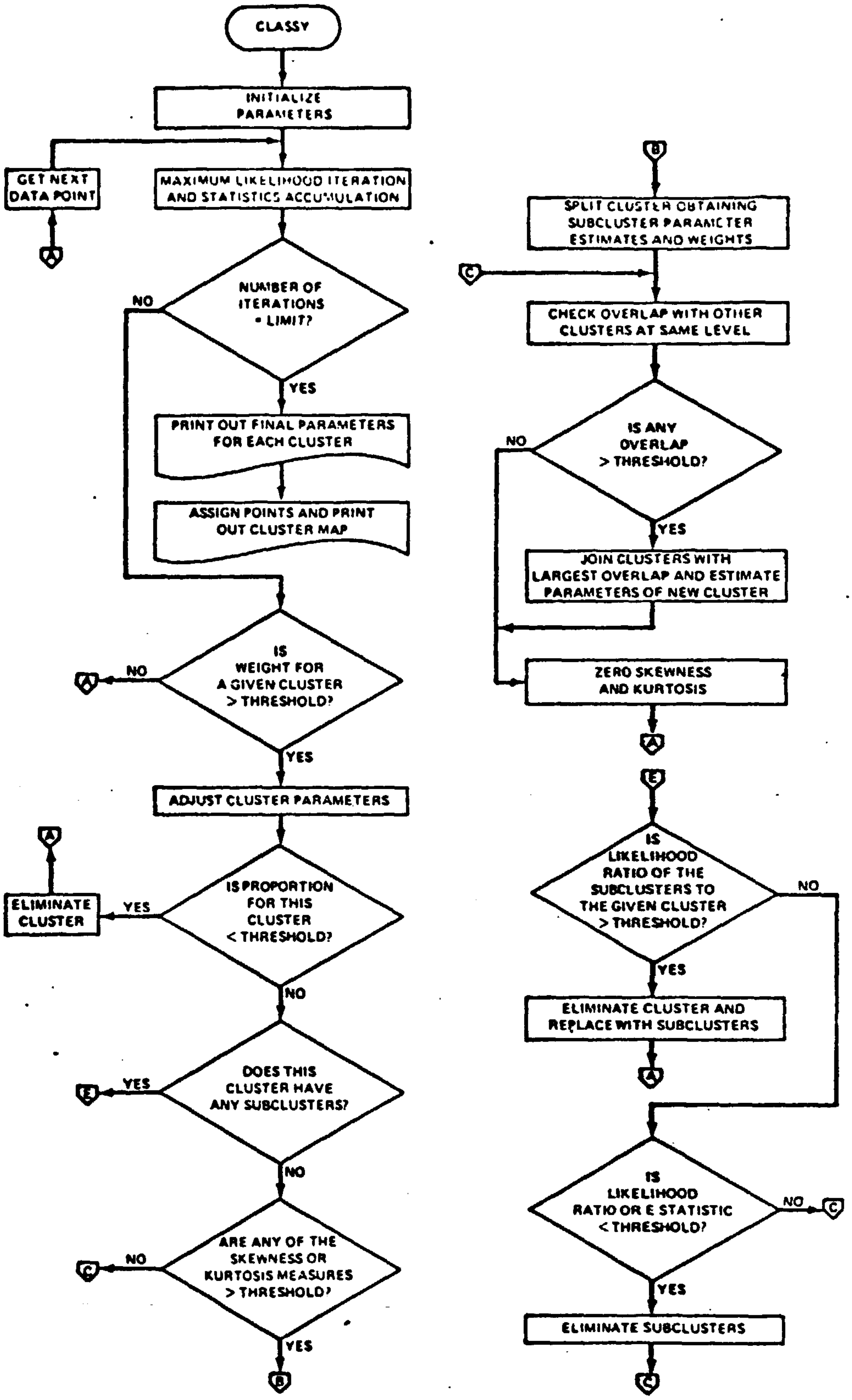
CLASSY
INITIALIZE PARAMETERS
GET NEXT DATA POINT
A
MAXIMUM LIKELIHOOD ITERATION AND STATISTICS ACCUMULATION
NUMBER OF ITERATIONS = LIMIT?
NO
YES
PRINT OUT FINAL PARAMETERS FOR EACH CLUSTER
ASSIGN POINTS AND PRINT OUT CLUSTER MAP
IS WEIGHT FOR A GIVEN CLUSTER > THRESHOLD?
NO
A
YES
ADJUST CLUSTER PARAMETERS
A
ELIMINATE CLUSTER
YES
IS PROPORTION FOR THIS CLUSTER < THRESHOLD?
NO
DOES THIS CLUSTER HAVE ANY SUBCLUSTERS?
YES
E
NO
ARE ANY OF THE SKEWNESS OR KURTOSIS MEASURES > THRESHOLD?
NO
C
YES
B
B
SPLIT CLUSTER OBTAINING SUBCLUSTER PARAMETER ESTIMATES AND WEIGHTS
C
CHECK OVERLAP WITH OTHER CLUSTERS AT SAME LEVEL
IS ANY OVERLAP > THRESHOLD?
NO
YES
JOIN CLUSTERS WITH LARGEST OVERLAP AND ESTIMATE PARAMETERS OF NEW CLUSTER
ZERO SKEWNESS AND KURTOSIS
A
E
IS LIKELIHOOD RATIO OF THE SUBCLUSTERS TO THE GIVEN CLUSTER > THRESHOLD?
NO
YES
ELIMINATE CLUSTER AND REPLACE WITH SUBCLUSTERS
A
IS LIKELIHOOD RATIO OR E STATISTIC < THRESHOLD?
NO
C
YES
ELIMINATE SUBCLUSTERS
C

FIGURE 1.—Flow diagram for the CLASSY algorithm.

Data Sets

Two different data sets were used in the comparative evaluation of CLASSY and ISOCLS. The first was a set of Landsat acquisitions of four different LACIE segments. Each LACIE segment is 196 picture elements (pixels) per line by 117 lines and corresponds to a 5- by 6-nautical-mile area on the ground. The second data set was a group of four different simulated acquisitions of a simulated LACIE segment. Each of these data sets is described separately in the following paragraphs.

The four LACIE segments were selected on the basis of the availability of ground truth at regularly spaced pixels in the image and the provision of a representative sampling of LACIE segments in terms of field structure and the proportion of wheat present. Once the segments had been chosen, the acquisition that had the greatest separability, as measured by the Bhattacharyya distance, was selected. The Bhattacharyya distance was computed between wheat and nonwheat classes where the class statistics were obtained from ground-truth fields. The segment number and location, the acquisition date with the largest separability, and the ground-truth percentages of wheat and small grains for each segment are given in table I.

TABLE I.—Description of LACIE Sample Segments

Segment	*Location*	*Acquisition date*	*Ground truth, percent wheat*	*Ground truth, percent small grains*
1181	Kansas	Mar. 10, 1976	23.4	29.0
1988	Kansas	Nov. 8, 1975	33.0	33.0
1961	Kansas	July 18, 1976	8.2	8.2
1965	North Dakota	Aug. 8, 1976	41.6	47.0

The simulated data set consisted of four simulated Landsat acquisitions, each 196 pixels by 117 lines. This data set was generated by IBM for the Mission Planning and Analysis Division at the Johnson Space Center (ref. 11). Each "acquisition" was obtained first by specifying the mean vector and covariance matrix for 10 different classes. The class statistics for each class were specified so as to simulate the LACIE data for two wheat classes (W_1 and W_2), two barley classes (B_1 and B_2), two classes of grass (G_1 and G_2), two stubble classes (S_1 and S_2), and two classes of fallow (F_1 and F_2). The statistics for these classes were actually obtained from Landsat data representing an agricultural area in Hill County, Montana. Once the statistics for a given class were specified, independent samples were generated from a four-dimensional multivariate normal distribution having those statistics. These samples were then placed in rectangular fields arranged over the simulated segment. This process was repeated for each class and for each of the four acquisitions. The arrangement of the simulated fields over the segment was the same for each acquisition. The pattern of the simulated fields is given in figure 2.

W_1	G_2	B_1	S_1	W_2	S_2	W_2	W_1	G_1	B_2
F_2	W_2	G_1	W_1	S_1	S_2	G_2	B_2	W_1	B_1
W_1	G_1	S_2	G_2	S_1	W_2	B_2	W_2	B_1	F_1
G_2	S_1	W_2	B_1	S_2	W_1	W_2	G_1	F_1	B_2
W_2	W_1	G_1	B_1	W_1	S_1	G_2	S_2	B_2	W_2

FIGURE 2.—Distribution of classes in simulated segment.

Evaluation Method and Procedures

CLASSY was evaluated using a comparative analysis method in which the clustering results of CLASSY were compared with those of ISOCLS using the ground truth as a reference. The evaluation procedure consisted of two steps.

1. The CLASSY and ISOCLS algorithms were applied to each segment in each data set. CLASSY was run for three complete iterations through all the data in each segment. ISOCLS was run in the nearest neighbor mode with 40 ground-truth pixels as start-

ing vectors. In this mode, ISOCLS merely assigns pixels to the nearest starting vector measured in terms of L-1 distance rather than operating iteratively. This mode was chosen for ISOCLS because this was the manner in which the algorithm was currently being used in the LACIE project.

2. The clusters in the line printer map produced by each algorithm were analyzed by first recording the cluster symbol and the corresponding ground-truth label (either wheat or nonwheat) for each pixel where ground truth was available. These results were tabulated, so that the number of ground-truth wheat pixels and ground-truth nonwheat pixels falling in each cluster was known. The clusters were then labeled wheat or nonwheat by majority rule.

A measure of the accuracy of each clustering algorithm in separating wheat from nonwheat (or a measure of the overall purity of the wheat and nonwheat clusters) was computed by estimating the probability of correct classification (PCC) for the labeled clusters. This probability is given by

$$PCC = \sum_{i=1}^{m_1} P\left(O_i|O\right) P(O) + \sum_{i=1}^{m_2} P\left(W_i|W\right) P(W) \qquad (28)$$

where m_1 is the number of clusters labeled "other"; m_2 is the number of clusters labeled wheat; $P(O_i|O)$ is the probability that a pixel falls in the ith "other" cluster, given that it is other than wheat; $P(W_i|W)$ is the probability that a pixel falls in the ith wheat cluster, given that it is wheat; $P(W)$ is the a priori probability that a pixel is wheat; and $P(O)$ is the a priori probability that a pixel is other than wheat. Empirical proportions were used to estimate these probabilities and a priori values, resulting in the following estimate:

$$\hat{PCC} = \frac{1}{N_T}\sum_{i=1}^{m_1} N_{O_i|O} + \sum_{i=1}^{m_2} N_{W_i|W} \qquad (29)$$

where N_T is the total number of ground-truth pixels, $N_{O_i|O}$ is the number of ground-truth "other" pixels falling in the ith "other" cluster, and $N_{W_i|W}$ is the number of ground-truth wheat pixels falling in the ith wheat cluster. It is noteworthy that, to obtain an accurate estimate of PCC using equation (29), it is necessary that several ground-truth pixels fall in each cluster. Specifically, if there are clusters which have only one or two ground-truth grid-intersection pixels, the estimate of PCC will be biased on the high side.

As a part of the analysis, the proportion of wheat was also estimated for the labeled clusters and compared to the ground-truth value. The equation used for this estimate is

$$\hat{P}(W) = \frac{1}{N_T}\sum_{i=1}^{m_2} N_{W_i} \qquad (30)$$

where N_{W_i} is the total number of ground-truth pixels (wheat and other) falling in the ith wheat cluster.

Estimates computed using equations (29) and (30) were obtained for each algorithm as applied to both the real and simulated data sets.

Results

The results of these computations are given in tables II through XI. Tables II, III, V, and VI compare CLASSY and ISOCLS results for the LACIE segments examined; the corresponding results for simulated segment data are given in tables VII through XI.

Table II compares the number of clusters and the PCC estimates for ISOCLS ($\hat{PCC}_I$) and for CLASSY ($\hat{PCC}_c$) as a result of clustering each of the four LACIE segments examined using both methods. The PCC estimates for CLASSY are, on the average, about 4 percentage points lower than those for ISOCLS. However, since the version of ISOCLS used

TABLE II.—Comparison of the Number of Clusters and the Estimated Probability of Correct Classification Using Single-Pass Segment Data

Segment	*ISOCLS*		*CLASSY*		$\hat{PCC}_c - \hat{PCC}_I$
	No. of clusters	$\hat{PCC}_I$	*No. of clusters*	$\hat{PCC}_c$	
1181	40	0.8410	7	0.8052	−0.0358
1988	40	.8070	8	.7661	−.0409
1961	40	.9236	11	.9028	−.0208
1965	40	.7419	9	.6774	−.0645
Average	40	.8284	8.75	.7875	−.0405

generates a factor of 4 to 6 times as many clusters as CLASSY, many of the ISOCLS clusters contain only one or two ground-truth grid-intersection points. As discussed in the preceding section, this means that the PCC estimates for ISOCLS will be biased high relative to CLASSY. In addition, each ISOCLS cluster typically contains one ground-truth point used as a starting vector for that cluster. Since the label of these starting vectors almost always agrees with the cluster label, this amounts to a further high bias in the PCC estimates for ISOCLS. In the light of this bias in favor of ISOCLS and the economy represented by the greatly reduced number of CLASSY clusters, CLASSY compares very favorably to ISOCLS.

The LACIE segments used in this study contained varying amounts of wheat. The ground-truth percentages of wheat $P(W)$ and small grains $P(SG)$ are given in table III. The estimate of the proportion of wheat computed using the ground-truth grid-intersection dots $\hat{P}_D(W)$ is also included. An estimate of the proportion of wheat in the whole scene determined from the clusters labeled wheat can be obtained using equation (30). The wheat proportion estimates resulting from applying this equation to the CLASSY results $\hat{P}_c(W)$ and ISOCLS results $\hat{P}_I(W)$ are also given in table III. Comparing these percentages to the ground-truth wheat proportions shows that, with the exception of segment 1965, the wheat proportion estimates are about 4 to 6 percent higher than the ground-truth wheat proportion values. These slightly high estimates may be due to the fact that, even though only wheat ground-truth dots were used to label clusters, labeled wheat clusters may reasonably be assumed to include some small grains. The last column in table III shows that the ISOCLS estimate was closer to the ground-truth wheat proportion for two segments and the CLASSY estimate was closer for the other two segments.

The imagery for segment 1965 was examined in detail because the wheat proportion estimates for both CLASSY and ISOCLS deviated considerably from the ground truth and the PCC estimates for both algorithms were correspondingly low for this segment. This segment contained numerous small strip fields. Typically, small-field regions accentuate misregistration problems, and such appears to be the case for this segment. The misregistration of the ground-truth reference acquisition relative to the acquisition clustered reduced PCC values and distorted the proportion of wheat estimates for both algorithms.

To obtain an idea about the relative performance of CLASSY and ISOCLS when applied to multitemporal data, four-channel "green" images were formed for each segment by applying the Kauth (ref. 12) transformation to each of four acquisitions for a given segment and then selecting the green number from each acquisition. (It was necessary to reduce the 16-dimensional data to 4 dimensions since CLASSY is limited to 4 dimensions at the present time.) Table IV lists the four acquisitions used for each segment. The results of comparing the PCC values and the wheat proportion estimates for the two algorithms are given in tables V and VI, respectively. Comparing table V and table II shows that the PCC values for both algorithms remained about the same for segments 1181 and 1961 and that they increased significantly for segments 1988 and 1965. The average difference between the CLASSY and ISOCLS PCC values remained about 4 percent. However, the CLASSY PCC equaled the ISOCLS PCC for segment 1988, and the difference was very small for segment 1961. The last column of table VI

TABLE III.—Comparison of Wheat Proportion Estimates for Labeled Clusters Using Single-Pass Segment Data

Segment	*Ground truth*		*Ground-truth dots* $\hat{P}_D(W)$	*ISOCLS* $\hat{P}_I(W)$	*CLASSY* $\hat{P}_c(W)$	$D_I = \hat{P}_I(W) - \hat{P}(W)$	$D_c = \hat{P}_c(W) - \hat{P}(W)$	$\lvert D_I\rvert - \lvert D_c\rvert$
	P(W)	P(SG)						
1181	0.234	0.290	0.333	0.287	0.303	0.053	0.069	−0.016
1988	.330	.330	.322	.397	.287	.067	−.043	.024
1961	.082	.082	.097	.042	.069	−.040	−.013	.027
1965	.416	.470	.516	.526	.645	.110	.229	−.119
Average	.266	.293	.317	.313	.326	.047	.061	−.021

TABLE IV.—Acquisitions Used in Creating Four-Channel Green Images

Segment	*Acquisitions*
1181	Mar. 10, 1976 Apr. 16, 1976 May 3, 1976 July 14, 1976
1988	Oct. 20, 1975 May 6, 1976 June 12, 1976 Sept. 28, 1976
1961	Aug. 15, 1975 June 12, 1976 Aug. 23, 1976 Sept. 10, 1976
1965	May 11, 1976 July 21, 1976 Aug. 8, 1976 Sept. 14, 1976

shows that, when the four-channel green images were used, the wheat proportion estimates from the CLASSY clusters were closer to the ground-truth values than were the ISOCLS estimates in every case.

Tables VII and VIII are analogous to tables II and III, except that they give the results for the single-pass simulated data. The column labeled maximum likelihood PCC (PCC_{ML}) gives the overall PCC when using standard maximum likelihood classification where the statistics for each class were computed from fields in the simulated image given the class label for each field. Note that the PCC estimates for CLASSY were higher than those for ISOCLS in two of the four passes. In fact, on pass 2, where the separability was greatest, the PCC for CLASSY equaled the maximum likelihood PCC. On the average, the PCC for CLASSY was 1.4 percent higher than that for ISOCLS.

TABLE V.—Comparison of the Number of Clusters and the Estimated Probability of Correct Classification Using the Four-Channel Green Image Data

Segment	*ISOCLS*		*CLASSY*		$P\hat{C}C_c - P\hat{C}C_I$
	No. of clusters	$P\hat{C}C_I$	*No. of clusters*	$P\hat{C}C_c$	
1181	40	0.8667	4	0.8000	−0.0667
1988	40	.9357	16	.9357	0
1961	40	.9167	23	.9097	−.0070
1965	40	.8065	13	.7290	−.0775
Average	40	.8814	14	.8436	−.0378

The proportion estimate computed from the labeled clusters is given in table VIII. Again, the estimate from CLASSY was closer to the true value in two of the four passes. However, the average individual ISOCLS estimate was about 2 percent closer to the true value.

The results for the simulated data using band 1 from each of the four passes are given in table IX. Band 1 was selected arbitrarily to assess the use of multitemporal data. Note that the PCC estimate for CLASSY was 1.0, meaning that none of the CLASSY clusters contained a mixture of wheat and nonwheat grid-intersection pixels.

Using the simulated data makes it possible to identify a cluster with a certain class in the data by determining which class contributes the majority of pixels to the cluster. After such an identification, the generating statistics for the class may be compared with the cluster statistics produced by CLASSY. Table X presents the results of such a comparison for

TABLE VI.—Comparison of Wheat Proportion Estimates for Labeled Clusters Using Four-Channel Green Image Data

Segment	*Ground truth*		*ISOCLS* $\hat{P}_I(W)$	*CLASSY* $\hat{P}_c(W)$	$D_I = \hat{P}_I(W) - \hat{P}(W)$	$D_c = \hat{P}_c(W) - \hat{P}(W)$	$\lvert D_I \rvert - \lvert D_c \rvert$
	P(W)	P(SG)					
1181	0.234	0.230	0.292	0.241	0.058	0.007	0.051
1988	.330	.330	.316	.342	−.014	.012	.002
1961	.082	.082	.066	.069	−.016	−.013	.003
1965	.416	.470	.625	.565	.209	.149	.060
Average	.266	.293	.325	.304	.059	.039	.029

TABLE VII.—Comparison of the Number of Clusters and the Estimated Probability of Correct Classification Using Single-Pass Simulated Data

Pass	PCC_M	*ISOCLS*		*CLASSY*		$PCC_M - P\hat{C}C_I$	$PCC_M - P\hat{C}C_c$	$P\hat{C}C_c - P\hat{C}C_I$
		No. of clusters	$P\hat{C}C_I$	*No. of clusters*	$P\hat{C}C_c$			
1	0.935	40	0.9139	5	0.9043	0.021	0.030	−0.0096
2	.986	40	.9713	5	.9857	.015	.000	.0144
3	.970	40	.9761	8	.9522	−.006	.018	−.0239
4	.928	40	.8852	7	.9187	.043	.009	.0335
Average	.955	40	.9366	6.25	.9402	.018	.014	.0144

TABLE VIII.—Comparison of the Wheat Proportion Estimates for Labeled Clusters Using Single-Pass Simulated Data

Pass	P(W)	$\hat{P}_I(W)$	$\hat{P}_c(W)$	$D_I = \hat{P}_I(W) - \hat{P}(W)$	$D_c = \hat{P}_c(W) - \hat{P}(W)$	$\lvert D_I\rvert - \lvert D_c\rvert$
1	0.3398	0.3301	0.2536	−0.0097	−0.0862	−0.0765
2	.3398	.3254	.3541	−.0144	.0143	.0001
3	.3398	.3636	.2917	.0238	−.0481	−.0243
4	.3398	.3254	.3349	−.0144	−.0049	.0095
Average	.3398	.3361	.3086	−.0147	−.0312	−.0228

the pass 2 simulated data, whereas table XI gives similar results for the clustering using band 1 from each of the four passes.

In the pass 2 CLASSY results, four of the five clusters could be clearly identified with one of the generating classes or distributions. A comparison of the mean vector and covariance matrices shows a remarkable correspondence between the CLASSY statistics and the generating statistics. Cluster 3 was almost equally divided between grass 1 and grass 2. Only the statistics for grass 1 are shown in the table. Similarly, cluster 2 was a mixture of stubble, fallow, and barley 2. The statistics for each of these classes are very similar for this pass. The statistics for stubble 1 are given as a representative example of that group of clas~es.

TABLE IX.—Probability of Correct Classification Using Multipass Simulated Data

ISOCLS		*CLASSY*		$P\hat{C}C_c - P\hat{C}C_I$
No. of clusters	$P\hat{C}C_I$	*No. of clusters*	$P\hat{C}C_c$	
40	0.9809	7	1.0000	0.0191

The data from band 1 of each of the four simulated passes had more separability; thus, CLASSY was able to distinguish more classes. The comparison of the generating statistics and the CLASSY statistics is presented in table XI. Only the variance terms from the multipass covariance matrix were available. Again, there is remarkable correspondence between the CLASSY statistics and the generating statistics.

CONCLUSIONS

The main conclusion of this paper is that the performance of the CLASSY clustering algorithm compares favorably with that of ISOCLS on both the real and simulated LACIE segment data. In terms of performance, these results were obtained despite the

TABLE X.—Comparison of Cluster Statistics for Pass 2 Simulated Data

Cluster number	*Identification*	*Generating statistics*					*CLASSY statistics*				
		Mean vector	*Covariance matrix*				*Mean vector*	*Covariance matrix*			
4	Wheat 1	20.36	0.91	1.21	0.34	−0.01	20.56	1.21	1.04	0.13	−0.19
		20.19	1.21	3.24	.24	−.65	20.67	1.04	2.87	−.10	−.95
		27.29	.34	.24	1.77	1.75	27.45	.13	−.10	1.84	1.76
		28.14	−.01	−.65	1.75	3.15	28.26	−.19	−.95	1.76	3.50
5	Wheat 2	18.55	0.82	0.69	−0.01	−0.47	18.76	1.08	0.80	−0.03	−0.50
		17.02	.69	1.11	−.48	−1.19	17.13	.80	1.54	−.47	−1.20
		26.35	−.01	−.48	1.23	1.41	26.36	−.03	−.47	1.46	1.50
		28.00	−.47	−1.19	1.41	3.25	27.97	−.50	−1.20	1.50	3.51
1	Barley 1	23.30	1.55	1.74	1.22	0.96	22.97	2.00	1.97	1.83	1.41
		25.80	1.74	3.16	1.52	1.12	25.45	1.97	3.59	2.36	1.77
		25.98	1.22	1.52	1.65	.91	25.27	1.83	2.36	2.92	1.84
		24.19	.96	1.12	.91	1.19	23.50	1.41	1.77	1.84	2.22
3	Grass 1 (grass 2)	20.83	1.31	2.07	0.54	0.11	20.71	1.15	1.48	0.55	0.22
		20.86	2.07	4.70	.91	−.29	20.54	1.48	4.10	1.01	.28
		23.37	.54	.91	1.10	.70	23.18	.55	1.01	1.40	.64
		22.50	.11	−.29	.70	1.23	22.52	.22	.28	.64	1.24
2	Stubble 1 (stubble 2, fallow, barley 2)	21.90	0.97	0.62	0.77	0.69	22.40	0.96	0.44	0.31	0.22
		23.64	.64	1.12	.70	.66	24.43	.44	1.17	.38	.29
		24.22	.77	.70	1.51	1.40	24.18	.31	.38	1.41	.92
		23.12	.69	.66	1.40	2.31	22.77	.22	.29	.92	1.68

fact that CLASSY reduces the number of clusters by a factor of 4 to 6 as compared to ISOCLS. This performance indicates that CLASSY is indeed approximating the empirical mixture density rather than just breaking up the data space into small homogeneous areas as does ISOCLS. This conclusion is further substantiated by noting the high degree of correspondence between the CLASSY cluster statistics and the generating statistics of classes in the simulated data. When data from band 1 of each of the 4 simulated acquisitions was clustered using CLASSY, 5 of the 10 classes were very accurately identified. The remaining classes, whose statistics were very close together, were broken into two reasonable groups. It appears, therefore, that the CLASSY algorithm may well provide a solution to the fundamental problem of clustering—the determination of the inherent number of classes in the data.

REFERENCES

1. MacDonald, R. B.; Hall, F. G.; and Erb, R. B.: The Use of LANDSAT Data in a Large Area Crop Inventory Experiment (LACIE). Proceedings of the Symposium on Machine Processing of Remotely Sensed Data, Purdue Univ. (W. Lafayette, Ind.), 1975, pp. 1B-1 through 1B-23.

2. Kan, E. P. F.: The JSC Clustering Program ISOCLS and Its Applications. LEC-0483, Lockheed Electronics Co., Houston, Tex., July 1973.

3. Turley, R. C.: Algorithm Simulation, Test, and Evaluation Program (ASTEP)—Users Guide and Software Documentation. NASA Johnson Space Center, Houston, Tex., Apr. 1978, pp. 1-22.

4. Ball, G. H.; and Hall, D. J.: A Clustering Technique for Summarizing Multivariate Data. Behavioral Science, vol. 12, Mar. 1967, pp. 153-155.

TABLE XI.—Comparison of Cluster Statistics for Band 1 for Each of Four Passes of the Simulated Data

Cluster number	Identification	Generating statistics					CLASSY statistics				
		Mean vector	Covariance matrix				Mean vector	Covariance matrix			
5	Wheat 1	26.93	1.06				26.84	1.27	0.69	1.42	1.61
		20.36		0.91			20.27	.69	1.21	1.25	1.62
		17.39			2.15		17.22	1.42	1.25	2.32	2.65
		17.27				3.30	17.02	1.61	1.62	2.65	3.49
2	Wheat 2	25.79	1.03				25.90	1.22	0.94	0.78	0.98
		18.55		0.82			18.76	.94	1.23	.78	.87
		16.85			0.47		16.88	.78	.78	.85	.67
		18.12				1.76	17.97	.98	.87	.67	1.80
4	Barley 1	28.41	2.16				28.40	2.30	1.56	3.03	2.18
		23.30		4.86			22.71	1.56	1.81	2.69	2.17
		22.01			4.15		22.56	3.03	2.69	5.33	3.80
		17.01				4.47	17.44	2.18	2.17	3.86	3.58
3	Barley 2	28.23	1.33				28.40	1.63	−0.08	1.79	1.05
		22.78		0.77			22.71	−.08	.79	−.40	−.09
		22.37			1.88		22.56	1.79	−.40	2.54	1.23
		17.34				1.61	17.44	1.05	−.09	1.23	1.86
1	Grass 1 (grass 2, stubble 1)	25.67	1.81				25.82	2.69	0.87	1.76	2.17
		20.83		1.31			21.20	.87	1.39	.74	.98
		20.10			1.80		20.35	1.76	.74	1.71	1.65
		20.60				1.62	20.72	2.17	.98	1.65	2.43
6	Fallow 1	24.59	0.67				24.68	0.75	0.38	0.42	0.48
		22.48		0.52			22.45	.38	.72	.68	.09
		23.22			0.90		23.21	.42	.68	1.06	.04
		21.56				0.66	21.67	.48	.09	.04	.75
7	Stubble 2 (fallow 2)	24.33	1.17				24.34	1.31	0.38	−0.01	−0.14
		22.21		0.67			22.25	.38	.86	.09	−.15
		22.69			0.74		22.70	−.01	.09	1.01	.84
		28.63				1.04	28.63	−.14	−.15	.84	1.35

5. MacQueen, J.: Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, L. M. LeCam and J. Neyman, eds., Univ. of California Press, 1967, pp. 281-297.

6. Duda, R. O.; and Hart, P. E.: Pattern Classification and Scene Analysis. John Wiley & Sons (New York), 1973.

7. Wolfe, J. H.: Pattern Clustering by Multivariate Mixture Analysis. Multivariate Behavioral Research, vol. 5, 1970, pp. 329-350.

8. Quirein, J. A.; and Trichel, M. C.: Acreage Estimation, Feature Selection, and Signature Extension Dependent Upon the Maximum Likelihood Decision Rule. Proceedings of the Symposium on Machine Processing of Remotely Sensed Data, Purdue Univ. (W. Lafayette, Ind.), 1975, pp. 2A-26 through 2A-39.

9. Day, N. E.: Estimating the Components of a Mixture of Normal Distributions. Biometrika, vol. 56, 1969, pp. 463-474.

10. Hasselblad, V.: Estimating the Parameters for a Mixture of Normal Distributions. Technometrics, vol. 8, 1966, pp. 431-466.

11. Oliver, R. E.: Description of Simulated LACIE Segments 1851-1854. IBM Information Systems Technical Report, Nov. 18, 1975.

12. Kauth, R. J.; and Thomas, G. S.: The Tasselled Cap—A Graphic Description of the Spectral-Temporal Development of Agricultural Crops as Seen by Landsat. Proceedings of the Symposium on Machine Processing of Remotely Sensed Data, Purdue Univ. (W. Lafayette, Ind.), 1976, pp. 4B-41 through 4B-51.

Appendix

The equation used to obtain iterative estimates of the a priori class probabilities or proportions, a_i, is derived beginning with equation (9), which is repeated here in a slightly more expanded form.

$$a_i = \frac{1}{N}\sum_{k=1}^{N}\frac{a_i p_i\left(x_k|\mu_k,\Sigma_i\right)}{p\left(x_k\right)} \qquad \text{(A1)}$$

where

$$p\left(x_k\right) = \sum_{j=1}^{m} a_j p_j\left(x_k|\mu_j,\Sigma_j\right)$$

Since a_i does not depend on k, a_i may be canceled from both sides of the equation to obtain

$$1 = \frac{1}{N}\sum_{k=1}^{N}\frac{p_{ik}}{p_k} \qquad \text{(A2)}$$

where, for convenience, the functional notation has been simplified.

Now,

$$0 = \frac{1}{N}\sum_{k=1}^{N}\frac{p_{ik} - p_k}{p_k} \qquad \text{(A3)}$$

But

$$\begin{aligned} p_k &= a_i p_{ik} + \sum_{\substack{j=1 \\ j\neq i}}^{m} a_j p_{jk} \\ &= a_i p_{ik} + \left(1 - a_i\right) q_{ik} \end{aligned} \qquad \text{(A4)}$$

Here, define

$$q_{ik} = \begin{cases} \displaystyle\sum_{\substack{j=1 \\ j\neq i}}^{m}\left(\frac{a_j}{1 - a_i}\right) p_{jk}, a_i \neq 1 \\ 0, \text{otherwise} \end{cases} \qquad \text{(A5)}$$

So,

$$\begin{aligned} 0 &= \frac{1}{N}\sum_{k=1}^{N}\frac{p_{ik} - a_i p_{ik} - \left(1 - a_i\right) q_{ik}}{p_k} \\ &= \frac{1}{N}\sum_{k=1}^{N}\frac{\left(1 - a_i\right) p_{ik} - \left(1 - a_i\right) q_{ik}}{p_k} \end{aligned} \qquad \text{(A6)}$$

$$0 = \sum_{k=1}^{N} \frac{p_{ik} - q_{ik}}{p_k} \tag{A7}$$

assuming $a_i \neq 1$. Breaking this sum up into those terms which are positive and those which are negative results in

$$0 = \sum_{p_{ik} > q_{ik}} \frac{p_{ik} - q_{ik}}{p_k} + \sum_{p_{ik} < q_{ik}} \frac{p_{ik} - q_{ik}}{p_k} \tag{A8}$$

Now, a_i is reintroduced as follows:

$$0 = \frac{1}{a_i}\left[a_i \sum_{p_{ik} > q_{ik}} \frac{p_{ik} - q_{ik}}{p_k}\right] + \frac{1}{(1 - a_i)}\left[(1 - a_i)\left(\sum_{p_{ik} < q_{ik}} \frac{p_{ik} - q_{ik}}{p_k}\right)\right] \tag{A9}$$

If we now solve for the a_i's which are outside the square brackets in terms of the a_i's, p_i's, and q_i's inside the square brackets, the following is obtained.

$$a_i = \frac{a_i \sum\limits_{p_{ik} > q_{ik}} \frac{p_{ik} - q_{ik}}{p_k}}{a_i \sum\limits_{p_{ik} > q_{ik}} \frac{p_{ik} - q_{ik}}{p_k} - (1 - a_i) \sum\limits_{p_{ik} < q_{ik}} \frac{p_{ik} - q_{ik}}{p_k}}$$

$$= \frac{a_i \sum\limits_{p_{ik} > q_{ik}} \frac{p_{ik} - q_{ik}}{p_k}}{- \sum\limits_{p_{ik} < q_{ik}} \frac{p_{ik} - q_{ik}}{p}}$$

$$= \frac{a_i \sum\limits_{p_{ik} > q_{ik}} \frac{p_{ik} - q_{ik}}{p_k}}{N - \sum\limits_{p_{ik} > q_{ik}} \frac{q_{ik}}{p_k} - \sum\limits_{p_{ik} < q_{ik}} \frac{p_{ik}}{p_k}} \tag{A10}$$

This is the iterative equation used to obtain proportion estimates in CLASSY.

Equation (A10) may also be put into a form illustrating the nature of the update term to obtain

$$a_i = a_i + \frac{a_i\,(1 - a_i) \sum\limits_{k=1}^{N} \frac{p_{ik} - q_{ik}}{p_k}}{N - \sum\limits_{p_{ik} > q_{ik}} \frac{q_{ik}}{p_k} - \sum\limits_{p_{ik} < q_{ik}} \frac{p_{ik}}{p_k}} \tag{A11}$$

This equation illustrates that direct functional iteration using equation (A10) amounts to adding a correction term given by

$$\frac{a_i\,(1 - a_i) \sum\limits_{k=1}^{N} \frac{p_{ik} - q_{ik}}{p_k}}{N - \sum\limits_{p_{ik} > q_{ik}} \frac{q_{ik}}{p_k} - \sum\limits_{p_{ik} < q_{ik}} \frac{p_{ik}}{p_k}}$$

to the old value of a_i in order to obtain the new value of a_i.

As a way of comparing the iterative equation for proportion estimates used in CLASSY (eq. (A10)) to the standard maximum likelihood iterative equation (eq. (A1)), one may rework the standard equation so that the nature of the update term is apparent. Using equation (A6), one obtains

$$1 = 1 + \frac{(1 - a_i)}{N} \sum_{k=1}^{N} \frac{p_{ik} - q_{ik}}{p_k} \tag{A12a}$$

or

$$a_i = a_i + \frac{a_i\,(1 - a_i)}{N} \sum_{k=1}^{N} \frac{p_{ik} - q_{ik}}{p_k} \tag{A12b}$$

This equation reduces exactly to equation (A1).

A comparison of equations (A11) and (A12) shows that the difference is in the term N versus

$$N - \sum_{p_{ik} > q_{ik}} \frac{q_{ik}}{p_k} - \sum_{p_{ik} < q_{ik}} \frac{p_{ik}}{p_k}$$

Thus, the iterative equation used in CLASSY (eq. (A11)) will amplify the correction for a_i if there are a significant number of points such that $0 < p_{ik} < 1$ and $0 < q_i < 1$. This corresponds to the case where cluster i is a "mixed" cluster; that is, there is a significant amount of overlap between cluster i and other clusters. Since it is precisely these "mixed" clusters for which the standard iterative equation (eq. (A1) or (A12)) converges slowly, the iterative equation used for proportions in CLASSY (eq. (A10) or (A11)) should converge more readily.