EVALUATION OF CLUSTERING TECHNIQUES FOR CROP AREA ESTIMATION USING REMOTELY SENSED DATA

Michael E. Bellow, Martin Ozga, USDA/NASS*
Michael E. Bellow, South Bldg., Room 4168, Washington, D.C. 20250

KEY WORDS: CLASSY, ISODATA, classification, signature

1. INTRODUCTION

Finding a suitable clustering algorithm has long been a problem when processing remotely sensed digital data from satellites for crop area estimation. In the PEDITOR software system used by the National Agricultural Statistics Service (NASS), the data presented to a clustering algorithm are usually assumed to represent a single crop or ground cover type. This assumption is justified since NASS has always had a large amount of ground information available from the enumerative surveys used to do area estimation without remotely sensed data. The addition of satellite data is intended to improve the quality of the estimates by providing additional inputs into a regression estimator [1]. The data, usually from Landsat or SPOT satellites, consist of a number of pixels (picture elements). Each pixel represents an area on the ground and has several channels, each representing a scaled value of reflectance in a particular spectral band. The scaling is between 0 and 255 so that each channel takes one byte of storage. Typical numbers of channels are seven for the Landsat Thematic Mapper (TM) and three for the French SPOT multispectral scanner. Multitemporal data are often used to help distinguish crops that may have similar spectral characteristics in a single scene. The multitemporal data consist of two scenes over the same area but from different dates, thus containing twice as much data as the single (unitemporal) scene. In order to reduce the computational burden, a subset of the available channels is often selected and used for processing.

The spectral characteristics of a given crop or cover type are known as its signature. The signature is affected by various factors both internal and external to the crop in question. The internal factors relate to the crop's species and variety, as well as its stage of development at the time a satellite image is taken. The external factors refer to atmospheric conditions present at the overpass times, as well as subsequent processing operations done on the data before delivery to the user.

* The authors thank James Mergerson for helpful comments and suggestions.

The task of a clustering algorithm is to generate clusters representing distinct categories in the data, but also representing the entire data set. The approach of having each data set represent a specific crop or cover type is known as supervised clustering.

Two clustering algorithms, known as ISODATA [2] and CLASSY [3], have been studied. Each is well known, but has been subject to some changes made locally. The main basis for comparison is the quality of estimates obtained from the same data through clustering with both programs. Clustering effectiveness was evaluated via statistical measures computed by programs in PEDITOR [4,5], the software system used for all computations described here.

First, the two algorithms as implemented at NASS will be described, followed by the results of the comparison.

2. CLUSTERING METHODOLOGY

ISODATA and CLASSY both attempt to find a collection of clusters that represent the input data set. A cluster is represented by a mean value vector and a variance-covariance matrix. The means and covariances for all clusters are used to perform a maximum likelihood classification over a large area, typically an entire satellite scene. The results of the classification are used as inputs to a regression estimator to obtain the crop area estimates.

With both algorithms, the user can adjust various parameters that govern the clustering. A short initialization step is followed by a series of iterations, each having two steps. The first step involves splitting and merging of clusters, while the second consists of a series of smaller iterations that do cluster adjustments. The entire procedure stops if no more splits or merges are possible, or if some criterion involving user specified parameters is satisfied.

The CLASSY algorithm was developed specifically for use with spectral data from satellites. It is based on the assumption of a multivariate normal mixture model for the data. The program attempts to estimate the number of components of the mixture via a sequence of

hypothesis tests using a likelihood ratio criterion. The parameters of each component are estimated using the iterative fixed point equations that result from a maximum likelihood formulation. By contrast, ISODATA makes no assumptions about the distribution of the data.

In ISODATA, the user selects an initial and minimum number of clusters. During the initialization step, the entire data set is viewed as a hypercube. The initial mean value vectors of the clusters are taken at evenly spaced intervals along the diagonal of this hypercube. During the cluster adjustment step that follows, each pixel is assigned to the cluster nearest to it in terms of ordinary Euclidean distance. The cluster mean value vectors and variance-covariance matrices are recomputed from those pixels assigned to the cluster. This procedure continues until a measure of convergence, the percentage of pixels that do not change clusters between two iterations, exceeds a threshold value. The threshold is selected by the user and usually falls between 98 and 100 percent. The ISODATA algorithm exits when either no more splits or merges are possible, the minimum number of clusters specified by the user has been reached, or the optional maximum number of iterations has been reached.

Splitting is only attempted on clusters for which a dispersion measure, the largest eigenvalue of the variance-covariance matrix, is larger than a threshold. The two clusters to be formed by the split are initialized using the ISODATA initialization step, but only for pixels assigned to the cluster to be split. The cluster adjustment is then performed on those same pixels, for the two new clusters only. Once convergence is attained, the new clusters are tested for validity. The split is retained only if the dispersion measures of the two new clusters are less than that of the original cluster by a user specified ratio, and if the number of pixels assigned to each new cluster is larger than a specified minimum value. If the split is retained, the original cluster is discarded in favor of the new ones. Merging occurs when the Swain-Fu distance [6], a measure of separation between clusters, is less than a threshold. The merge is a straightforward union of the two clusters, weighted by the number of pixels in each.

CLASSY is more complex than ISODATA in many ways. A key feature of CLASSY is that, unlike ISODATA, no pixel is completely assigned to any cluster. Instead, there is a probability generated for each pixel belonging to each cluster, known as the weight of the pixel relative to the cluster. Each cluster also has a weight, defined to be the mean of the weights of all pixels relative to that cluster. The initialization step of CLASSY simply computes a single cluster based on the entire data set, setting all pixel weights relative to that cluster to 1, and therefore the cluster weight also to 1. Unless there is very little variability in the data, new clusters should be generated by splits over the next few iterations. The cluster adjustment step consists of a series of maximum likelihood iterations in which the weights as well as the mean value vectors and variance-covariance matrices are adjusted. The test of convergence is maximum percentage of change in cluster weights according to a parameter set by the user, usually between two and five percent. The entire CLASSY algorithm exits if no splits or merges are possible, or if the number of large iterations exceeds a value set by the user.

When clusters are split, the original cluster is not necessarily discarded. Instead, a tree of clusters is formed in which the split clusters are the children of the original cluster. The tree may be pruned during a merge, as will be seen. As the tree grows and shrinks with split and merge decisions, the number of children for any particular cluster may become larger than two, so the tree is not a binary tree. However, if a particular cluster has only one child, that child will be deleted during a periodic tree cleanup.

Clusters are eligible for splitting if the skew or kurtosis values of the variance-covariance matrices exceed a threshold. Only end node clusters are eligible for splitting. The initial mean value vectors and variance-covariance matrices for the split clusters are obtained in a manner similar to the split routine in ISODATA, but using those pixels having the largest weights for the cluster being split. Cluster adjustment is then performed via the maximum likelihood iterations, again using only those pixels with the largest weights for the original cluster.

Merging takes place on clusters based on a similarity value derived from the mean value vectors, variance-covariance matrices, and cluster weights. It occurs when this similarity value is higher than a threshold. Since it is possible for a particular cluster to have a similarity value higher than the threshold with more than one other cluster, the similarity values of all pairs of clusters exceeding the threshold are sorted and processed in descending order, being careful not to process any cluster more than once. The effect of the merge varies depending upon the relative positions of the two clusters in the tree, leading to cluster deletion or actual merging. The merge is a straightforward combination of the mean value vectors and variance-covariance matrices, weighted by their cluster weights.

CLASSY has been rewritten to better fit into the PEDITOR environment. In addition, two major changes have been made from the earlier version. The first is that all pixels are processed in all iterations, rather than the random sample used in the old version. The second is that the cluster split routine has been changed, making it similar to the one used in ISODATA.

3. INPUT DATA

The two clustering algorithms were compared using Landsat Thematic Mapper (TM) data from 1988 for regions of western Iowa and eastern Arkansas. The crops of interest in Iowa were corn and soybeans, while in Arkansas they were cotton, rice, and soybeans. The Iowa data were unitemporal, with a satellite overpass date of July 25. The Arkansas data were multitemporal, with overpass dates of May 17 and August 5. For Iowa, all seven TM channels were used. For Arkansas, channels 2 through 5 from both the early and late season scenes were used, resulting in an eight dimensional data set.

The following is a brief description of how the input data for the clustering algorithms were obtained. Each spring, NASS conducts the June Agricultural Survey (JAS), a national sample survey that uses both an area frame and a list frame. The area frame part of the survey involves a stratification of each state's area into land cover classes. Within each stratum, the land is further subdivided into sampling units known as segments, usually one square mile. Enumerators visit a random sample of segments from each stratum and collect data on crops planted in specific fields, as well as location of features such as roads, woods, and water.

For this study, the satellite scenes covering the Iowa and Arkansas regions were registered to a map base so that pixels corresponding in location to the JAS fields could be identified. Pixels whose ground data contained more than one cover type were removed and the remaining pixels were placed in special files, called packed files. All covers containing fewer than five percent of the total number of pixels in the area covered by the sample segments were combined into a single packed file. A clipping algorithm based on principal components [7] was used to remove outlier pixels. The most prevalent cover types in the Iowa region were corn (45% of the sample area), soybeans (31%), and permanent pasture (7%). The main covers in the Arkansas region were soybeans (32%), rice (17%), idle cropland (16%), woods (13%), and cotton (10%).

4. PERFORMANCE MEASURES

The test runs done for each cover using ISODATA and CLASSY were compared using three internal and four external clustering criteria. The internal criteria were among the best from a large number studied by Milligan and Cooper [8,9], using Monte-Carlo methods. They measure an algorithm's ability to minimize within-cluster variability while maximizing separation between clusters. The internal measures are as follows:

1. Calinski-Harabasz index:

$$C.H. = \frac{(m-c)\sum_{i=1}^{c}[m_i|\bar{z}_i-\bar{z}|^2]}{(c-1)\sum_{i=1}^{c}\sum_{j=1}^{n_i}|z_{ij}-\bar{z}_i|^2]}$$

2. B/W index:

$$B/W = [d_b/f_b]/[d_w/f_w]$$

3. Point-biserial correlation coefficient:

$$P.B. = [d_b-d_w][f_w f_b/f_d^2]^{1/2}/s_d$$

where:

- m = number of pixels in data set
- c = number of clusters formed
- m_i = number of pixels in cluster i (i=1,..,c)
- z_{ij} = vector of spectral values for cluster i, pixel j (j=1,...,m_i)
- $\bar{z}_i$ = mean vector of pixels in cluster i
- $\bar{z}$ = mean vector of all pixels in data set
- d_b = sum of pairwise between-cluster distances between pixels
- d_w = sum of pairwise within-cluster distances between pixels
- f_b = number of between-cluster pixel pairs
- f_w = number of within-cluster pixel pairs
- f_d = total number of pixel pairs (= m(m-1)/2)
- s_d = standard deviation of all pairwise distances

The pairwise distances referred to are Euclidean distances. The three criteria are positive measures of clustering effectiveness. C.H. is an adjusted ratio between sums of squared distances, analogous to an F-statistic. B/W is the ratio between the mean between-cluster and mean within-cluster pairwise distances. The point-biserial coefficient is a measure of correlation between the set of pairwise distances and a variable taking the values 0 or 1 according to whether or not two pixels are from the same cluster.

The external performance measures are related to operations done on the data after clustering, namely classification and regression estimation. Following each set of clustering runs, all pixels within the sample segments were classified to a cover. The categories and discriminant functions formed via clustering were used by the classification program. Prior probabilities for the categories, computed from information on relative acreage of the covers in the region of interest, were used to adjust the discriminant functions.

Two measures, percent correct and commission error (C.E.), are direct indicators of classification accuracy. Percent correct is the percent of pixels reported for a given cover type that were classified to that cover. Commission error is the percent of those pixels classified to a cover that were reported to a different cover. Another index commonly used is overall percent correct, the percent of all pixels in the data set classified to their reported cover type.

Within the sample area for specified strata, the NASS crop area estimation procedure uses regression methodology to relate classified pixel counts to the ground reference data. For this study, only one stratum per state was used. The counts of pixels within each sample segment classified to a given cover were regressed against the corresponding acreage values from the JAS enumeration. A first order regression model was used, generating standard least squares parameter estimates.

In operational remote sensing, the regression equations within strata are used to compute large scale (region level) crop area estimates, based on classification of all pixels in the scenes covering the region. Proration is used to estimate crop acreages in areas where it is not feasible to use remotely sensed data (e.g. cloud covered areas). The large scale regression estimates can be compared with the direct expansion estimates computed using only JAS survey data.

The key criterion used by NASS to evaluate remote sensing estimation accuracy is the regression coefficient of determination:

$$R^2 = \frac{[\sum_{j=1}^{n} (x_j-\bar{x})(y_j-\bar{y})]^2}{\sum_{j=1}^{n} (x_j-\bar{x})^2 \sum_{j=1}^{n} (y_j-\bar{y})^2}$$

where:

n = number of segments
x_j = number of pixels classified to crop in segment j
y_j = reported acres of crop in segment j
$\bar{x}$ = mean pixels per segment classified to crop
$\bar{y}$ = mean acres per segment reported to crop

R^2 measures the goodness of fit of the regression equation. It is closely related to relative efficiency (R.E.), the ratio between the variances of the direct expansion and regression estimates.

5. RESULTS

For the Iowa data set, the crops of interest were corn and soybeans. The other two cover types used for clustering were permanent pasture and 'other' (all remaining covers combined). For Arkansas, the crops of interest were cotton, rice, and soybeans, and the additional covers were idle cropland, woods, and 'other'. Although performance measures were computed for all of these covers, estimation efficiency for the crops of interest is most important to NASS.

The two clustering programs were run on the same data sets, with the input parameters used being the default values. These defaults had been chosen previously after extensive testing showed that they gave the best performance among all sets of values tested. The sample sizes were 26 segments for Iowa and 22 for Arkansas.

Table 1 gives the computed values of seven clustering measures for each cover type tested. It is seen that ISODATA produced clusterings with higher values of the three internal criteria for all covers in each state. This indicates that ISODATA was more effective than CLASSY in producing compact, well defined clusters.

The main basis for selecting a clustering algorithm is the quality of the resulting area estimates, as measured by the regression coefficient of determination. Table 1 shows that in Iowa, ISODATA generated a higher value of R^2 than CLASSY for corn, permanent pasture, and 'other', but a lower value for soybeans. In Arkansas, R^2 was higher with ISODATA than CLASSY for five of the six cover types, with 'other' being the exception.

To assess whether ISODATA produced a significantly better regression fit than CLASSY, F-tests for equality of residual variances were performed on all cover types for both states. The residuals were assumed to be independent, identically distributed, and normal with mean

zero. The tests were one-sided with the alternative being that the variance of the regression residuals was smaller for ISODATA than for CLASSY. The test statistic F^* is equal to the ratio between the sums of squared residuals for ISODATA and CLASSY, respectively. Table 2 gives the value of F^*, degrees of freedom (same for numerator and denominator), and approximate p-value for each cover type. From the p-values, it is seen that at the 10 percent level, ISODATA resulted in a significantly smaller residual variance than CLASSY for two covers in Iowa (corn and 'other'), and two in Arkansas (cotton and soybeans). This represents three of the five crops of interest in the two states.

Tables 1 and 3 give the classification accuracy measures for the two data sets. Percent correct was higher with ISODATA than CLASSY for all four covers in Iowa and five of the six covers in Arkansas. The commission error was lower with ISODATA except for one Iowa cover and one Arkansas cover. ISODATA showed a higher overall percent correct than CLASSY in both states.

6. DISCUSSION

The results showed that ISODATA produced more compact, well-defined clusters than CLASSY, leading to overall better classification and estimation accuracy. However, the disparity in performance was not that great. The fact that, in a few cases, CLASSY gave higher values of certain performance measures than ISODATA is evidence that the algorithm may still be useful.

The current study represents a preliminary assessment of the performance of CLASSY, following the complete reworking of the algorithm. Further research could lead to refinements that would improve the clustering efficiency. ISODATA has been evaluated more thoroughly and is less likely to be modified in the near future.

The results presented here led to a decision to use only ISODATA in the near future. A longer term clustering strategy for future operational programs will be developed based on further research on the two algorithms. Possible areas for future investigation include the effect of the pixel data distribution on performance of the algorithms, the degree of improvement achieved by using multitemporal instead of unitemporal data, and the effect of data sampling on clustering effectiveness.

REFERENCES

[1] J.D. Allen, "A Look at the Remote Sensing Applications Program of the National Agricultural Statistics Service," Journal of Official Statistics, vol. 6, no. 4, pp. 393-409, 1990.

[2] G.H. Ball and D.J. Hall, "A Clustering Technique for Summarizing Multivariate Data," Behavioral Science, vol. 12, pp. 153-155, March 1967.

[3] R.K. Lennington and M.E. Rassbach, "CLASSY - An Adaptive Maximum Likelihood Clustering Algorithm" in Proceedings of the Ninth Annual Meeting of the Classification Society (North American Branch), Clemson, South Carolina, May 21-23, 1978.

[4] M. Ozga, "USDA/SRS Software of Landsat MSS-Based Crop Acreage Estimation", in Proceedings of the IGARSS '85 Symposium, Amherst, Massachusetts, October 7-9, 1985, pp. 762-772.

[5] G. Angelici, R. Slye, M. Ozga, and P. Ritter, "PEDITOR - A Portable Image Processing System," in Proceedings of the IGARSS '86 Symposium, Zurich, Switzerland, Sept. 8-11, 1986, pp. 265-269.

[6] P.H. Swain, "Pattern Recognition: A Basis for Remote Sensing Data Analysis," Information Note 111572 (1973), Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana.

[7] S.B. Winings, M. Ozga, and J. Stakenborg, "The Effects of the Application of Smoothing and Orthogonal Transforms to SPOT and TM Data on Regression Based Crop Acreage Estimates," in Proceedings of the IGARSS '90 Symposium, College Park, Maryland, May 20-24, 1990, pp. 647-649.

[8] G.W. Milligan, "A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis," Psychometrika, vol. 46, no. 2, pp. 187-199, June 1981.

[9] G.W. Milligan and M.C. Cooper, "An Examination of Procedures for Determining the Number of Clusters in a Data Set," Psychometrika, vol. 50, no. 2, pp. 159-179, June 1985.

Table 1: Clustering Performance Measures

IOWA

Cover	Method	No. Cats.	C.H.	B/W	P.B.	R^2	R.E.	% Corr.	% C.E.
Corn	ISODATA	16	7,355.5	2.75	.324	.915	10.83	85.03	20.80
	CLASSY	17	1,341.1	1.71	.203	.851	6.18	82.62	25.45
Soybeans	ISODATA	11	9,819.3	3.00	.415	.908	9.99	82.93	20.23
	CLASSY	15	2,888.2	1.89	.274	.927	12.64	76.60	17.53
Perm. Pasture	ISODATA	3	2,360.0	1.83	.515	.729	3.39	49.99	53.84
	CLASSY	14	562.9	1.35	.162	.692	2.99	40.62	57.45
Other	ISODATA	5	6,489.0	2.28	.477	.879	7.60	51.70	36.11
	CLASSY	21	1,006.9	2.06	.237	.688	2.95	50.58	44.72

ARKANSAS

Cover	Method	No. Cats.	C.H.	B/W	P.B.	R^2	R.E.	% Corr.	% C.E.
Cotton	ISODATA	4	3,866.4	2.06	.517	.989	84.52	87.75	11.75
	CLASSY	9	739.4	1.18	.106	.977	39.94	82.26	21.10
Rice	ISODATA	7	12,904.7	3.17	.540	.937	14.46	87.96	15.12
	CLASSY	18	2,500.5	1.58	.226	.935	14.02	82.28	22.33
Soybeans	ISODATA	10	10,738.2	2.68	.506	.844	5.80	81.29	22.61
	CLASSY	15	4,680.8	1.80	.351	.716	3.18	66.23	29.16
Idle Cropland	ISODATA	6	5,565.4	2.22	.518	.776	4.04	67.02	39.43
	CLASSY	10	2,355.2	1.54	.291	.723	3.27	58.39	44.50
Woods	ISODATA	7	7,502.4	4.25	.688	.749	3.61	76.87	23.58
	CLASSY	18	1,915.0	2.20	.197	.737	3.44	72.06	25.22
Other	ISODATA	5	4,484.9	2.36	.718	.443	1.62	36.95	44.90
	CLASSY	19	1,484.2	1.94	.266	.581	2.16	56.85	44.35

Table 2: Results of F-tests on Regression Residuals

State	Cover	F^*	df	p-value
Iowa	corn	.570	24	.09
	soybeans	1.266	24	>.5
	permanent pasture	.881	24	.4
	other	.388	24	.013
Arkansas	cotton	.472	20	.05
	rice	.970	20	.5
	soybeans	.549	20	.095
	idle cropland	.809	20	.4
	woods	.952	20	.5
	other	1.329	20	>.5

Table 3: Overall Percent Correct

State	Method	Overall % Correct
Iowa	ISODATA	73.74
	CLASSY	70.06
Arkansas	ISODATA	75.02
	CLASSY	69.09