AUTOMATIC CATEGORIZATION OF CLUSTERS IN UNSUPERVISED CLASSIFICATION

Sunpyo HONG, Kiyonari FUKUE, Haruhisa SHIMODA, Toshibumi SAKATA

Tokai University Research and Information Center
2-28-4 Tomigaya, Shibuya-ku, Tokyo 151, JAPAN

ABSTRACT:

A cluster categorization method is necessary when an unsupervised classification is used for remote sensing image classification. It is desirable that this method is performed automatically, because manual categorization is a highly time consuming process.

In this paper, several automatic determination methods were proposed and evaluated. They are 1) maximum number method, which assigns the target cluster to the category which occupies the largest area of that cluster; 2) maximum percentage method, which assigns the target cluster to the category which shows the maximum percentage within the category in that cluster; 3) minimum distance method, which assigns the target cluster to the category having minimum distance with that cluster; 4) element ratio matching method, which assigns the local region to the category having the most similar element ratio of that region. From the results of experiments, it was certified that the result by the minimum distance method was almost the same as the result made by a human operator.

Key Words: unsupervised classification, post processing, categorization, clustering.

1. INTRODUCTION

With the launch of second generation high resolution sensors like LANDSAT TM and SPOT HRV, clustering method has been revaluated recently. However, the main problem of clustering for practical use is that clustering is an unsupervised classification. That is, clusters generated by clustering are defined in feature vector space, not in image data. Therefore, in order to use the classified result for a meaningful reference map, it is necessary to determine the relation of clusters and categories, and to label the classified result with the categories.

Conventionally, this relation has been determined mainly by interpretation of an operator. However, this process is time consuming and is not objective.

The purpose of this research is to try several methods of automatic categorization and find out the most useful method. In this paper, 4 methods have been examined.

2. PROBLEMS OF CONVENTIONAL METHOD

In this method, each classified cluster is overlaid with the target image data on the display, and that cluster is interpreted by an operator to determine the category. Therefore, it can be thought that the obtained result is natural and reliable.

However, since everything is determined by an operator in this method, there are many problems as follows.

(1) The result depends on the skill of an operator.
(2) Objective and quantitative evaluation is difficult.
(3) It is time consuming when the number of cluster is large or there are many small clusters.

3. AUTOMATIC CATEGORIZATION METHOD

To solve the above problems, several automatic categorization methods are considered as follows. In all methods, training category areas(TCA) are first extracted from the target image similar to supervised trainings.

(1) Maximum Number Method

In this method, the number of pixels in each TCA for each cluster is calculated. Then the category having the maximum number is assigned to that cluster.

(2) Maximum Percentage Method

In this method, for each cluster, the percentage(occupation rate) of that cluster in each TCA is calculated. Then the category having the maximum percentage is assigned to that cluster.

Fig. 1 shows a comparison of these two methods in a simple case. Suppose that cluster k is composed of three categories A, B and C. As shown in Fig. 1(a), category A occupies the largest area in cluster k and C occupies the minimum area. In the maximum number method, cluster k is always assigned to category A. However, this figure does not show the difference of areas of each category. Fig. 1(b) shows the case that the total area of each category is the same and (c) shows the case that the total area of each category is different. As shown from this figure, categories which occupy small areas in the image tends to be neglected in the maximum number method.
On the contrary, small area categories

are treated favorably in the maximum percentage method as shown in Fig. 1.

(3) Minimum Distance Method

In this method, first, TCA is overlaid with the target image and the centroid of ecah category is calculated. Secondly, the distance between each category and each cluster is calculated. Then the category having the minimum distance is assigned to that cluster. For the distance, Euclidian distance was used in this experiment.

In the case of the maximum number method and the maximum percentage method, the result is dependent upon the size and location of training area. In the minimum distance method, it needs less number of pixels for TCA compared to other 3 methods but needs caution on spectral featute of TCA.

(4) Element Ratio Matching Method

In this method, based on the idea that the category can be represented by the set of clusters, the relation of categories and cluster is determined by the ratio of clusters which compose a category. For defining the element ratio, area is necessary. Therefore, the process is performed by local region as follows.

At first, TCA is overlaid with the result of clustering, which is classified by the clusters, and the element ratio of each category is calculated. Secondly, for each local region that is fixed on the result of clustering, element ratio is calculated and matched with the element ratio of each category precalculated. Then the category having the most similar element ratio of concerned region is assigned to that region.

4. EXPERIMENTS AND RESULT

4.1 FLOW OF EXPERIMENTS

In order to evaluate the proposed methods described in chapter 3, following LANDSAT TM data was used in the experiment. Sagami River basin was seleted for target area to pereform the quantitative evaluation. This area includes the test site area which landcover is already investigated and categorized in 52 categories.

sensor	: LANDSAT TM
date	: Nov. 4, 1984
path-row	: 107-35
area	: Sagami River basin in Japan
pixel size	: 25m X 25m
image size	: 512 X 480 pixels

At first, clusters were generated by a hierarchical clustering using Ward method. Since the image data in remote sensing is very large, usually clustering is performed with sampled data. In this experiment, 2500 samples(about 1% of entire image data) were used to generate 66 clusters. Based on the 66 clusters, the target image data was classified by a maximum likelihood method. Secondly, representative area of each category in the target image(training category area) was selected. 14 categories were selected as shown in Table 1. Finally, the relations of clusters with categories were determined by 4 methods described in chapter 3.

To evaluate quantitatively, classification accuracy was estimated based on the test site area. The classification accuracy was calculated over 5 major categories as shown in table 1 to adjust the selected 14 categories in target image to 52 test site categories.

4.2 Results of Experiments

(1) Maximum Number Method

Fig. 2 shows the result by this method, and Table 2 shows the classification accuracy estimated with the test site data. As expected, this result shows the tendency that clusters were labeled with the categories having large area such as urban.

(2) Maximum Percentage Method

Fig. 3 shows the result by this method, and Table 2 shows the classification accuracy estimated with the test site data. As expected, this result shows the tendency that clusters were labeled with the categories having small area such as other.

(3) Minimum Distance Method

Fig. 4 shows the result by this method, and Table 2 shows the classification accuracy estimated with the test site area. This result is very similar to the result(Fig. 2) by a human operator. In other words, a good result was obtained with roughly selected TCA.

(4) Element Ratio Matching Method

Fig. 5 shows the result by this method, and Table 2 shows the classification accuracy estimated with the test site area. In this method, 5 X 5 sized local region was used.

(5) Conventional Supervised Method

For the purpose of comparison, conventional supervised method was executed. In this method, clusters were categorized by a human operator. 66 clusters were labeled with 14 categories (Table 1) as shown in Fig. 6. Table 2 shows the classification accuracy by this method with the test site area.

5. CONCLUSIONS

Four categorization methods for clusters were evaluated by experiments using LANDSAT TM data. From the results of experiments, following conclusions were obtained.

(1) In the case of the maximum number method, the clusters are hardly labeled

with the small area categories. On the contrary, in the case of the maximum percentage method, the clusters are hardly labeled with large area categories. These are general problems for these two methods, because usually the area of each category is not the same.

(2) The classification accuracies(area weighted mean) were the hightest for the element ratio matching method but detail information such as river or road was lost as shown in Fig. 6 because that cateogorization was performed on each local region.

(3) The classification accuracies(simple mean) were the hightest for supervised method, second, the minimum distance method. the lowest for the elemene ratio matching method.

(4) From the view point of practical use, the maximum number method,the maximum percentage method and the element ratio matching method are easy to execute because they don't need the spectral feature, but the classification accuracies (simple mean) of three methods are lower than that of the minimum distance method.

(5) Among the methods which were evaluated, the minimum distance method showed best result. In this method, obtained result is almost the same with a supervised method. Theoretically, this method also needs less number of pixels for TCA compared to other 3 methods because the geometrical information of TCA is not used.

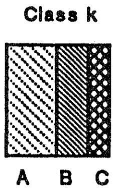
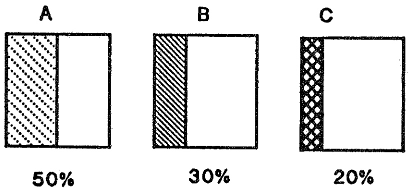
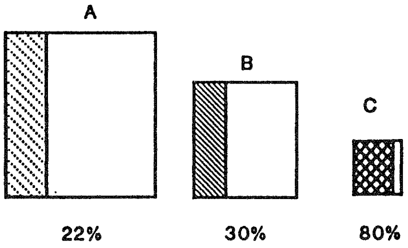
	category area and occupation rate of cluster k in that category area	categorized by maximum number method	categorized by max.percentage method
Class k  (a) cluster k, composed of category A,B,C	 (b) case of same size of category area	A	A
	 (c) case of different size of area	A	C

Fig. 1 Cluster and Categories

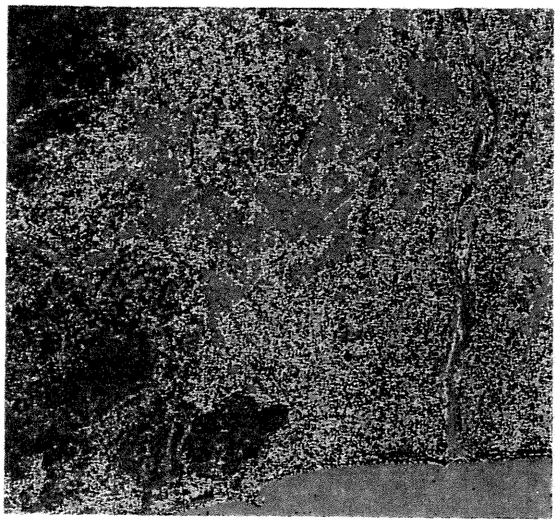


Fig. 2 Result by Maximum Number Method

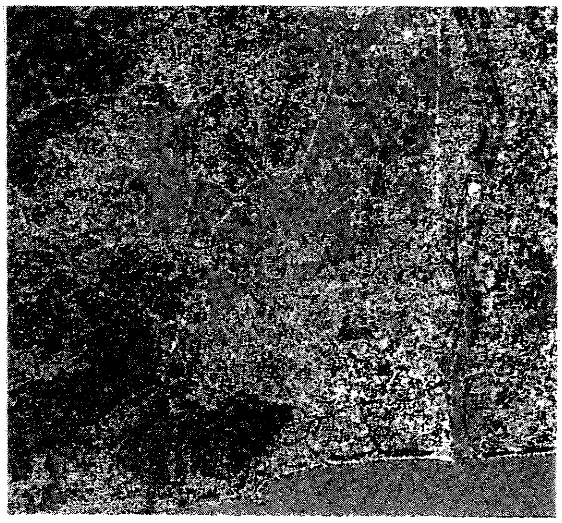


Fig. 3 Result by Maximum Percentage Method

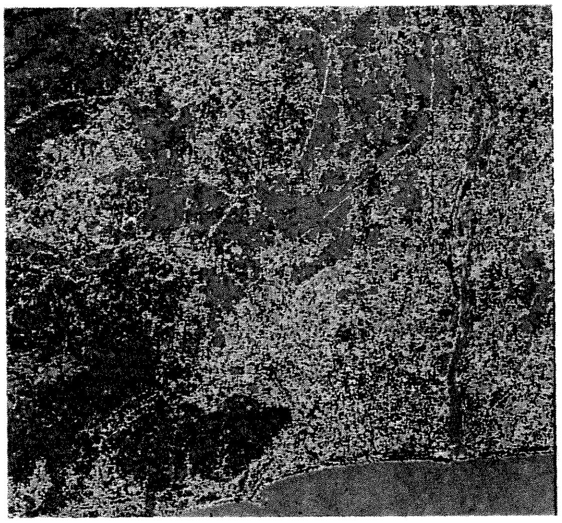


Fig. 4 Result by Minimum Distance Method

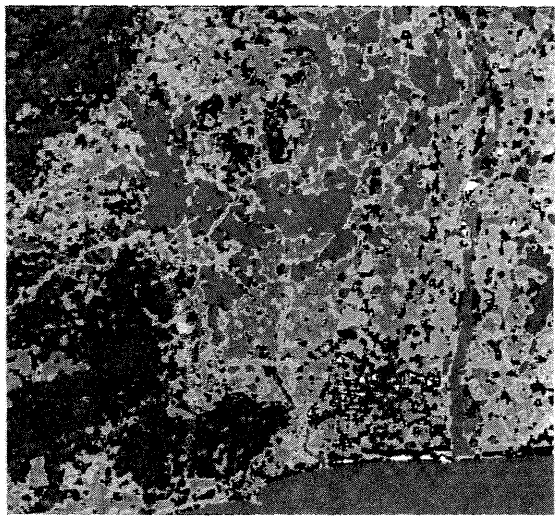


Fig. 5 Result by Element Ratio Matching

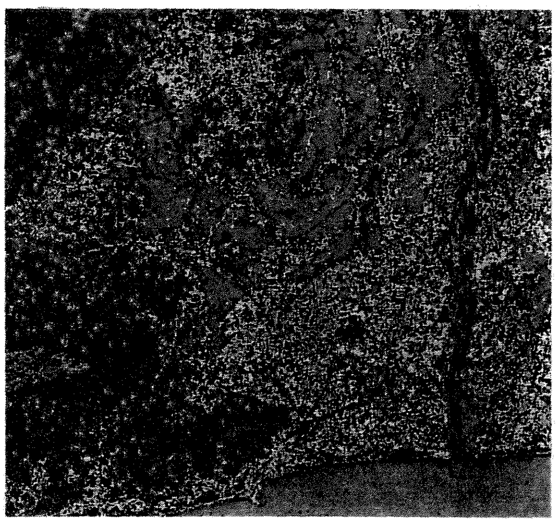


Fig. 6 Result by Supervised Method

Table 1 Categories and Clusters

categories			clusters			
Major	selected category	test site category	Supervised	Number	Percent.	Dist.
TREE	coniferous tree	coniferous tree,	3	7	5	4
	broad leaved tree	broad leaved tree,	9	4	4	3
	mixed tree	mixed tree, bamboo	4	3	5	7
	shadow	orchard, etc,	2	2	2	2
PADDY	paddy	paddy	10	11	10	8
URBAN	city area	concrete, factory,	6	4	2	3
	house area	building, railway,	6	11	13	11
	factory	house, urban, etc	7	8	8	7
WATER	sea	sea, river,	2	2	2	2
	river	pool, pond	1	1	1	2
OTHER	farm	farm, vinyl house,	1	6	6	6
	grass land	ground, grassland,	10	4	4	3
	ground	lawn, gravel,sand,	5	3	3	8
	sand area	ordered land, etc	0	0	1	0
5	14	44	66	66	66	66

Table 2 Classification Accuracies unit: %

CATEGORY	TREE	PADDY	URBAN	WATER	OTHER	Weighted	Simple
(occupation Rate)	(6.8%)	(9.6%)	(49.9%)	(5.0%)	(28.7%)	Mean	Mean
Max. Number	55.9	58.3	84.4	51.7	26.6	61.7	55.4
Max. Percentage	48.1	61.7	75.9	51.7	35.6	59.9	54.6
Min. Distance	47.5	48.9	85.6	67.8	32.7	63.4	55.5
Element Ratio	33.8	51.5	91.2	58.3	33.6	65.4	53.7
Supervised	46.2	56.3	78.0	67.8	46.4	64.2	58.9