Reprinted from

# Eleventh International Symposium

# Machine Processing of

# Remotely Sensed Data

with special emphasis on

## Quantifying Global Process:
## Models, Sensor Systems, and Analytical Methods

**June 25 - 27, 1985**

# Proceedings

Purdue University
The Laboratory for Applications of Remote Sensing
West Lafayette, Indiana 47907 USA

# A CLUSTERING ALGORITHM FOR REMOTE SENSING MULTISPECTRAL DATA

LI DAWEI

Zhanjiang Electronic Industrial Co.
Zhanjiang, Guangdong, China (PRC)

## I. ABSTRACT

This paper describes a clustering algorithm, named SORT, for classification of remote sensing multispectral data. Using a comb shape area as the training area, modifying the cluster table just after a new cluster found, ordering the cluster table and scanning the scene by an improved nearest neighbor method the good performance has been shown since SORT ran in the image processing system IRSA II of CAS.

## II. INTRODUCTION

Remote sensing multispectral data are widely available in geoscience study and its application recently. In China, a new image processing system IRSA II mainly using this sort of data has been completed at Chinese Acedemy of Sciences in 1984. A clustering program SORT by auther is working successfully with some other classification softwares in this system. A similar program has worked on computer NOVA 840 for several years(Li Dawei, 1981). The image processing system IRSA II is based on computer Eclipse S140 connected with COMTAL. Through terminals of S140 the hardware and software resource of COMTAL can be shared. So some revision was made during the transplantation of SORT onto the new computer environment.

The abundant experience in digital image precessing field for remote sinsing data has been accumulated for more than ten years(G.Nagy, 1972)(Li Dawei, 1981).

It is noticed that the pixels neighboring to each others usually belong to the same object or cluster. So image data can be divided into several stripes along lines for classification(G.Nagy and J.Tolaba, 1972). For saving calculation time the very similar clusters should be merged togather when they appear.

It is pointed that frequently the area boundaries appear not in every band. So decision of clusters with separate bands is more effective than with all of bands (A.C.Wacker and D.A.Landgrebe, 1970). Timesaving classification with not bad accuracy can be gotten in separate band calculation(G.Nagy, 1972)(J.N.Gupta, R.L. Kettig, D.A.Landgrebe and P.A.Wintz, 1973).

Not every cluster can form a multivariate Gaussian probability distribution in N-dimensional feature space. The method using a chain of several subclusters for representation of a cluster which has not an 'eye-pleasing' shape is often reasonable(E.P.Kan, W.A.Holley and H.D.Parker, Jr., 1973).

## III. ALGORITHM DESCRPTION

Program SORT is written with a clustering classification algorithm, which Searches an Object and Refreshes the Table immediately. 'Object' means a set of data that can be assigned into one cluster. The procedure can be divided into two steps, construction of cluster table and scanning the whole image.

Let $\overline{z_i}$ and $\overline{z_j}$ are two feature vectors,

$$\overline{z_i}=(z_{i1},z_{i2},\ldots,z_{in}) \quad (1)$$

and

$$\overline{z_j}=(z_{j1},z_{j2},\ldots,z_{jn}) \quad (2)$$

then the block distance between them

$$d_{ij}=\sum_{l=1}^{n}|z_{il}-z_{jl}| \quad (3)$$

is used as a measure of similarity. The two feature vectors can be assigned into one cluster if

$$d_{ij}<T \quad (4)$$

where T is a threshold value. Using the mean vector $\overline{M_j}=(m_{j1},m_{j2},\ldots,m_{jn})$ of a cluster $C_j$ for the vector $\overline{z_j}$ in formula (3) the distance from pixel $z_i$ to cluster $C_j$ is obtained. It is to say that the pixel $z_i$ belongs to the cluster $C_j$ only if the relation (4) happens. The measure of similarity between two clusters is given by

$$d_{ij}=\sum_{l=1}^{n}\frac{|m_{il}-m_{jl}|}{c_{il}+c_{jl}} \quad (5)$$

where $m_{il}$, $m_{jl}$ are the mean values and $c_{il}$, $c_{jl}$ are the standard deviations. In fact, formula (5) is equivalent to the divergence between two clusters $C_i$ and $C_j$(Li Dawei, 1981).

The procedure starts from an initial cluster table. The data are scanned line by line. When a vector assigned into the reject set of the cluster table is found, regarding this vector as the mean vector a new cluster might be searched out by formula (4), then it will be rectified iteratively in this data line. Then the cluster table is expanded by adding this new cluster to it. The new cluster may be merged into an old one in the table when the distance between them (5) is close enough, or neglected because the too small population. Searching period is continued until no any new cluster can be found, then scanning will be transfered to the next line, and so on. Through merge and chain, and ordering the clusters by the populations from bigger to smaller the table will be constructed over.

A cluster starting from an arbitrarily selected vector can be available as an initial cluster table.

The remote sensed data usually have high correlation among the neighboring pixels. So it is suitable to take a comb shape set of data lines for quickly constructing the table.

An improved nearest neighbor method is used for SORT. A vector very closed to a cluster center in the table will be assigned into it immediately, otherwise the comparison should be made throughout the whole table.

Because the extremely unbalanced populations of the clusters ordering the cluster table is very needed for rapid calculation.

## IV. PRACTICE

The image processing system IRSA II is user-friendly, so is the program SORT. The whole system is operated with multilevel menu, some prompts, defaults and options. Because the connection of computer Eclipse S140 with COMTAL the image data can be input from magnetic tape, disk or directly from the memory of COMTAL and the classifying results can be stored onto tape or disk, or in real-time displayed on the screen of COMTAL in a pseudocolor coding picture.

SORT has been compared with some other algorithms and image processing systems such as maximum likelihood classifier, ISODATA/ISOCLS(G.H.Ball and D.J.Hall, 1966) (E.P.Kan, W.A.Holley and H.D.Parker, Jr., 1973), ECHO(R.L.Kettig and D.A.Landgrebe, 1976) and (G.Nagy and J.Tolaba, 1972), and Model 575 by $I^2S$. And SORT has processed the data of Beijing, Yellow River Basin and Tibit for land-use applications perfectly.

## V. CONCLUSION

SORT has been proved as a successful classification algorithm and run in the image processing system IRSA II for a lot of geoscience applications in China. The comb shape trainning area, searching only the new objects different from those in the table and refreshing the table instantly, and the modified nearest neighbor method all benefit SORT with good classification accuracy and efficiency.

## VI. ACKNOWLEDGEMENT

The auther expresses his gratitude to Y. S. Cui and C.G. Zhu of Chinese Academy of Sciences for their encouragement and help in software development and preparation of the manuscript.

## REFERENCES

1. G.H.Ball and D.J.Hall, ISODATA, an Iterative Method of Multivariate Analysis and Pattern Classification, Proc. of the International Communication Conference, Philadelphia, 1966.

2. J.N.Gupta, R.L.Kettig, D.A.Landgrebe and P.A.Wintz, Machine Boundary Finding and Sample Classification of Remote Sensed Agriculture Data, Proc. IEEE, vol.61, 1973.

3. E.P.Kan, W.A.Holley and H.D.Parker, Jr., The JSC Clustering Program ISOCLS and Its Applications, Proc. 1973 Symp. on Machine Processing of Remote Sensing Data, LARS, Purdue University.

4. R.L.Kettig and D.A.Landgrebe, Classification of Multispectral Image Data by Extraction and Classification of Homogeneous Objects, IEEE Trans. Geosci. Electron., vol.GE-14, 1976.

5. Li Dawei, A Study of Classification Algorithms for Remote Sensing Data, unpublished graduate degree thesis, Gratuate School of CAS, 1981.

6. G.Nagy, Digital Image-processing Activeties in Remote Sensing for Earth Resources, Proc. IEEE, vol.60, 1972.

7. G.Nagy and J.Tolaba, Nonsupervised Crop Classification through Airborne Multispectral Observations, IBM J.Res.Develop., vol.16, no.2, 1972.

8. A.G.Wacker and D.A.Landgrebe, Boundaries in MSS Imaging by Clustering, Proc., 9th IEEE Symp. Adaptive Processor, 1970.

## AUTHOR BIOGRAPHICAL DATA

Li Dawei(David Lee) was born in Tientsin, China. He graduated from Chengdu Telecommunication Engineering Institute, China in 1965, majoring in mathematics in electronics. He received his M.S. deree in digital processing from Graduate School of Chinese Academy of Sciences, Peking, China.in 1981. He has engaged in remote sensing image pro-

cessing field as a research assistant at Chinese Academy of Sciences since then. In 1983 he studied and made pretty good research of data processing under several programs by Geological Survey and LARS of Purdue University as a visiting scientist in the United States. Now he is employed by Zhanjiang Electronic Industrial Company as an engineer in charge of computer business.

He is keen about developing microcomputer-based image processing systems.