Determination of the Initialization Number of Clusters in K-means Clustering Application Using Co-Occurrence Statistics Techniques for Multispectral Satellite Imagery

Kitti Koonsanit, Chuleerat Jaruskulchai, and Apisit Eiumnoh

***Abstract*—Nowadays, clustering is a popular tool for exploratory data analysis, such as K-means and Fuzzy C-mean. Automatic determination of the initialization number of clusters in K-means clustering application is often needed in advance as an input parameter to the algorithm. In this paper, a method has been developed to determine the initialization number of clusters in satellite image clustering application using a data mining algorithm based on the co-occurrence matrix technique. The proposed method was tested using data from unknown number of clusters with multispectral satellite image in Thailand. The results from the tests confirm the effectiveness of the proposed method in finding the initialization number of clusters and compared with isodata algorithm.**

***Index Terms*—Determination a number of clusters, number of cluster, K-mean**

I. Introduction

Clustering is a popular tool for data mining and exploratory data analysis. One of the major problems in cluster analysis is the determination of the number of clusters in unlabeled data, which is a basic input for most clustering algorithms. In this paper, we propose a new easy method for automatically estimating the number of clusters in unlabeled data set. Pixel clustering technique in a color image is a process of unsupervised classification of hundreds thousands or millions pixels on the basis of their colors. One of the most popular and fastest clustering techniques is the k-means technique. The results of the k-means technique depend on different factors such as a method of determination of initial cluster centers as shown in Fig. 1. Such sensitivity to initialization is an important disadvantage of the k-means technique. In this paper, a method has been developed to determine the initialization number of clusters in satellite image clustering application using a data mining algorithm based on the co-occurrence matrix technique. Therefore, automatic determination of the initialization number of clusters can greatly help with the unsupervised classification of satellite Image.

Manuscript received May 9, 2012; revised June 15, 2012. This work was supported by Thailand Graduate Institute of Science and Technology (TGIST) is gratefully acknowledged. The scholar student ID is TG-22-11-53-005D and the grant number is TGIST 01-53-005.

Kitti Koonsanit and Chuleerat Jaruskulchai are with Kasetsart University, Bangkok, Thailand (e-mail: sc431137@hotmail.com).

Apisit Eiumnoh is with National Center for Genetic Engineering and Biotechnology, Patumthani, Thailand.

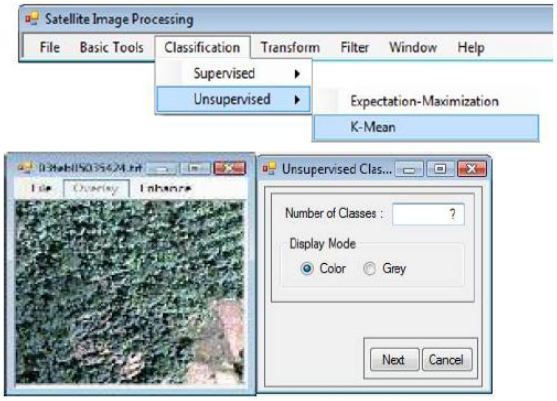


Fig. 1. Example k-mean algorithm in satellite application

II. Related Work

Many approaches to image classification have been proposed over the years [1]. Of these various methods, clustering is one of the simplest, and has been widely used in clustering of grey level images [2]-[4]. Techniques such as k-means[5], isodata[5], and fuzzy c-means[6], [7] have been around for quite a while, however, their application to color images has been limited. Although color images have increased dimensionality by requiring three bands such as red, green and blue, clustering techniques can be easily extended to cope with this. The k-means and fuzzy c-means algorithms require the number of clusters to be known beforehand.[8]-[11] In order to supply the information required by the aforementioned algorithms, the user must have some knowledge about the image, and this may not be the case. The new method is compatible with the k-means algorithm and it overcomes the limitation of having to indicate the number of clusters by co-occurrence matrix which is a apply technique in this proposed paper.

III. K-mean Method

The k-means method aims to minimize the sum of squared distances between all points and the cluster center. This procedure consists of the following steps, as described by Tou and Gonzalez [5].

TABLE I: Algorithm Finding A number of Clustering

Algorithm k- means
Input: k : the number of desired clusters Output: A set of k clusters
Processing
1. Select k objects from D as initial cluster centers 2. Form k clusters by assigning each object to its closest center 3. Recomputed the center of each cluster 4. Until centers do not change

Initially, the number of clusters must be known, or chosen, to be K say. The initial step is the choosing a set of K instances as centers of the clusters. Often chosen such that the points are mutually "farthest apart", in some way.

Next, the algorithm considers each instance and assigns it to the cluster which is closest. The cluster centroids are recalculated either after each instance assignment, or after the whole cycle of re-assignments. This process is iterated.

From table I, it is obvious in this description that the final clustering will depend on a method of initialization, i.e. the method of determining the initial cluster center. The classic version of the k-means uses a random selection of the initial centers. The aim of research is to find a fast initialization method that leads to high performance clustering and does not allow for formation of empty clusters. The new approach involves using co-occurrence statistics techniques as a method for the k-means initialization.

IV. The Proposed Algorithm

In this paper, we propose a new method for determination of the initialization number of clusters, which is based on co-occurrence matrix scheme. While a traditional co-occurrence matrix specifies only the transition within an image on horizontal and vertical directions. The proposed method can be used to automatically select a k range in multispectral satellite image as shown in Table II.

TABLE II: Algorithm Finding a Number of Clustering

Algorithm finding a number of clustering
Input: D : a data set containing n objects
Output: K : the number of desired clusters
Processing
1. Read D image
2. Transform D to G gray scale image
3. Read G gray scale image
4. Transform G to T co-occurrence matrix
5. Histogram = T.diagonal()
6. K = Finding Local Maximum (Histogram)

The proposed technique consists of two main steps. First, the co-occurrence matrix scheme is employed to automatically segment out the object region in an image. Then, the local maximum technique is used to count a number of regions, which a number of cluster, as shown in Fig. 2.

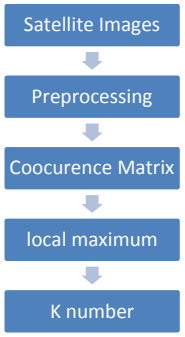


Fig. 2. Automatic determination of the initialization number of clusters

A. Co-Occurrence Matrix

Our test data are the satellite images which consist of different gray level as shown in Fig. 3 and Fig. 4.

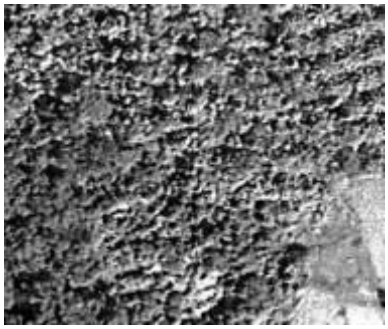


Fig. 3. Examples of satellite images

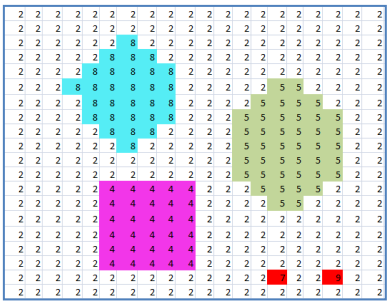


Fig. 4. Example of NIR band which is a data set containing n objects

Our definition of a co-occurrence matrix[12]-[14] is based on the idea that the neighboring pixels should affect region of clusters. Hence, we define a definition for a co-occurrence matrix by including the transition of the gray-scale value between the current pixel and adjacent pixel into our co-occurrence matrix illustrated in Fig. 5 and Fig. 6

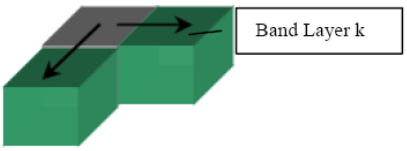


Fig. 5. right and bottom of pixel in a co-occurrence matrix

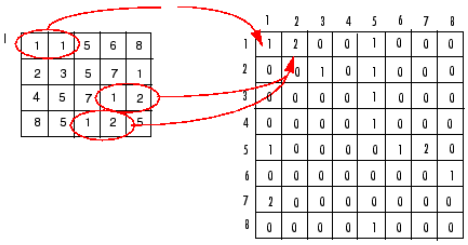


Fig. 6. Creating a co-occurrence matrix

Let F be the set of images. Each image has dimension of P×Q. Let t_{ij} be an element in a co-occurrence matrix depending upon the ways in which the gray level i follows gray level j

$$t_{ij} = \sum_{x=1}^{P} \sum_{y=1}^{Q} \delta \begin{cases} (F_k(x,y) = i) \text{ and } (F_k(x,y+1) = j) & \text{or} \\ (F_k(x,y) = i) \text{ and } (F_k(x+1,y) = j) \end{cases}$$

where $\delta = 1$, $\delta = 0$ *otherwise*. (1)

where F_k denotes the k^{th} band in the image set, F

If $s, 0 \le s \le L-1$ is a threshold. Then s can partition the co-occurrence matrix into 4 quadrants, namely A, B, C, and D shown in Fig.7.

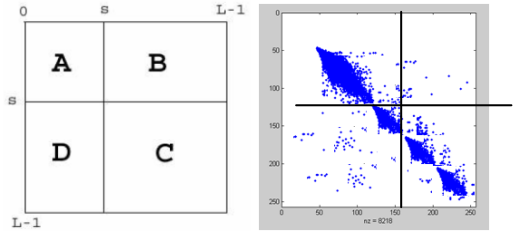


Fig.7. An example of blocking of co-occurrence matrix

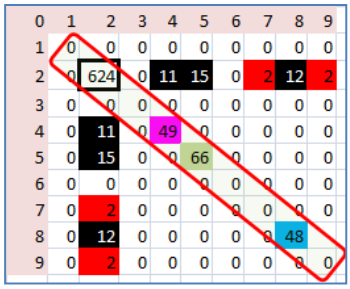


Fig. 8. An example of co-occurrence matrix of image from Fig. 4.

Since two of the quadrants shown in Figure.8, B and D, contain information about edges and noise alone, they are ignored in the calculation. Because the quadrants, which contain the object and the background, A and C, are considered to be independent distributions

B. Diagonal Matrix

The idea of proposed method is to select the results of co-occurrence matrix into a diagonal matrix. After threshold processing, the result of diagonal matrix was shown in Fig. 8. Diagonal matrix is used to show some clustered pixels. The gray level corresponding to local maximum which give the optimal number for object- classification in image as shown in Fig. 9.

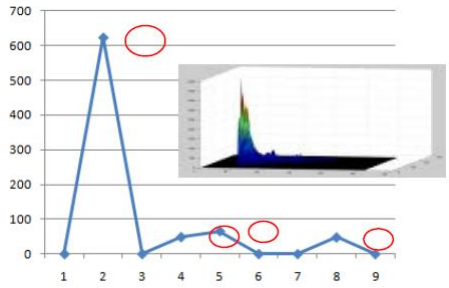


Fig. 9. Example histogram from diagonal of co-occurrence matrix (k=4) (1 background and 3 objects)

V. Experiment And Results

A. Dataset

We used the eight sets of raw data from different CCD Multi-spectrum images [15]. The dataset are obtained from small multi mission satellite project (SMMS), a department of Electrical Engineering, Kasetsart University. We would like to analyze data, which was registered in Thailand and thus try to determinate of the initialization number of clusters in interesting areas of Thailand.

B. Unsupervised Classification Method

The experiments performed in this paper use the simple K-mean from the Weka software package [16]. The simple K-Mean is the common unsupervised classification method used with remote sensing data. The effectiveness of the K-Mean depends on reasonably accurate estimation of the k cluster for each spectral class.

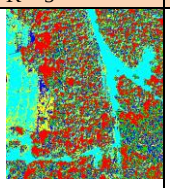
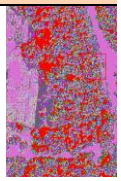
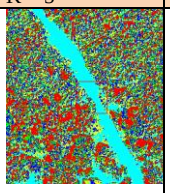
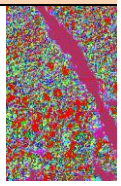
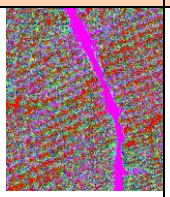
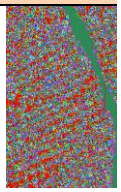
C. Experimental Result

Our experiment was tested with CCD Multi-spectrum images and shown in Table.III and Table.IV. The experiments demonstrate the robustness and effectiveness of the proposed algorithm.

In the paper this approach is successfully compared with isodata algorithm (iterative self-organizing data analysis technique) which is the top five frequently used unsupervised classification algorithms in remote sensing [17].

The isodata algorithm has some further refinements by splitting and merging of clusters. Clusters are merged if either the number of members (pixel) in a cluster is less than a certain threshold or if the centers of two clusters are closer than a certain threshold. Clusters are split into two different clusters if the cluster standard deviation exceeds a predefined value and the number of members (pixels) is twice the threshold for the minimum number of members.

TABLE III: Results From Experiment

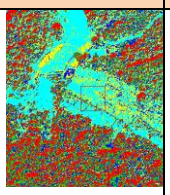
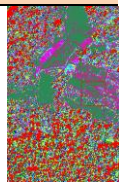
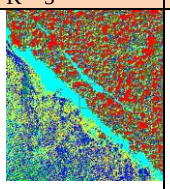
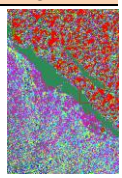
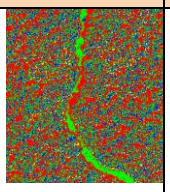
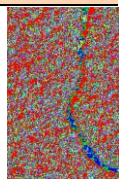
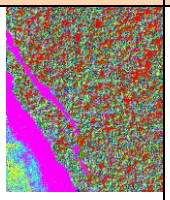
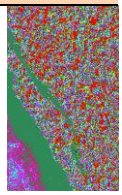
Set	Original Image	K Result from our experiment	K Result from isodata
I		K = 5	K=7
			
II		K = 5	K=7
			
III		K = 6	K=7
			

The isodata algorithm is similar to the k-means algorithm with the distinct difference that allows for min-max number of clusters while the k-means assumes that the number of clusters is known a priori.

We created using java with weka develop on netbean IDE version 6.8 and tested the proposed process on satellite image data such as small multi-mission satellite (SMMS), and the experimental results show that our proposed process can determine the initialization number of clusters in satellite image clustering effectively. The algorithm provides promising performance in determining of the initialization number of clusters in K-means clustering application by using co-occurrence statistics techniques for multispectral

satellite image.

TABLE IV: RESULTS FROM EXPERIMENT

Set	Original Image	K Result from our experiment	K Result from isodata
IV		K = 5	K=8
			
V		K = 5	K=8
			
VI		K=4	K=8
			
VII		K=6	K=8
			

From the experimental result, it was found that clustering using K solved by co-occurrence statistics techniques gives the nearest number of cluster with isodata. These selected K were used as input for K-mean clustering algorithms. Table III-IV shows the number of cluster obtained from various datasets for an example experiment. It can be noticed that the differences in clustering between K solved by isodata and K solved by co-occurrence statistics techniques are very closed.

The outcome of this research will be used in further steps for analysis tools in satellite image mining that helps K-mean method to visualize the satellite image such as natural resources and agricultural. A result of this research was developed to provide users have been processed to view and analyses the satellite image. We hope that it can be used as a tool and help develop research in satellite image data mining software in the future.

VI. CONCLUSION

Automatic determination of the initialization number of clusters in K-means clustering application by using co-occurrence statistics techniques for multispectral satellite image are presented in this paper. We define a definition for a co-occurrence matrix which can both preserve the structure within an image for clustering. The approach can be applied to automatically indicate an appropriate number of clusters range in satellite images. The algorithm provides promising performance in determining of the initialization number of clusters in K-means clustering application by using co-occurrence statistics techniques for multispectral satellite Image. Therefore, automatic determination of the initialization number of clusters can greatly help with the unsupervised classification of satellite Image.

ACKNOWLEDGMENT

The authors would like to thank a department of Electrical Engineering, Kasetsart University, Thailand for providing dataset on meteorological and Thailand Graduate Institute of Science and Technology (TGIST), a member of National Science and Technology Development Agency (NSTDA) for their financial support. The scholar student ID is TG-22-11-53-005D and the grant number is TGIST 01-53-005.

REFERENCES

[1] N. R. Pal and S. K. Pal, "A review on image segmentation techniques," *Pattern Recognition*, vol. 26, pp. 1277-1294, 1993.

[2] G. B. Coleman and H. C. Andrews, "Image segmentation by clustering," in *Proc. IEEE*, vol. 67, pp. 773-785, 1979.

[3] A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data," N*ew Jersey: Prentice Hall*, 1988.

[4] R. Nevatia, Image segmentation, In T.Y. Young and K.S. Fu (Eds.), *Handbook of Pattern Recognition and Image Processing*, Orlando: Academic Press, 1986.

[5] J. T. Tou and R. C. Gonzalez, "Pattern Recognition Principles," *Massachusetts: Addison-Wesley*, 1974.

[6] M. M. Trivedi and J. C. Bezdek, "Low-level segmentation of aerial images with fuzzy clustering," *IEEE Trans. Syst. Man. Cybern*., vol. SMC-16, pp. 589-598, 1986.

[7] Y. W. Lim and S. U. Lee, "On the color image segmentation algorithm based on the thresholding and the fuzzy c-means techniques," *Pattern Recognition*, vol. 23, pp. 935-952, 1990.

[8] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society*, 2001.

[9] M. M. T. Chiang and B. Mirkin, "Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads," *Journal of Classification*, vol. 27, no. 1, .2010

[10] M. M. T. Chiang and B. Mirkin, "Experiments for the number of clusters in K-means," in *Proceedings of the aritficial intelligence 13th Portuguese conference on Progress in artificial intelligence*, December 07-03, 2007.

[11] S. Ray and R. H. Turi, "Determination of number of clusters in K-means clustering and application in colour image segmentation," (invited paper) in N R Pal, A K De and J Das (eds), in *Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques (ICAPRDT'99)*, pp. 27-29, 1999. Narosa Publishing House, New Delhi, India, ISBN: 81-7319-347-9, pp. 137-143.

[12] N. R. Pal and S. K. Pal, "Entropic thresholding," *Signal processing*, vol. 16, pp. 97-108, 1989.

[13] T. Chanwimaluang and G. Fan, "An efficient algorithm for extraction of anatomical structures in retinal images," *ICIP 2003 Proceedings*, Sept 4-17, 2003

[14] K. Koonsanit, T. Chanwimaluang, D. Gansawat, S. Sotthivirat, W. Narkbuakaew, W. Areeprayolkij, P. Yampri, and W. Sinthupinyo, "Metal Artifact Removal on dental CT Scanned Image by Using Multi-layer Entropic Thresholding and Label Filtering Technique for 3-D Visualization of CT images," in *Proc. of International Conference on Biomedical Engineering : ICBME 2008.IFMBE Proceedings, 13th International Conference on Biomedical Engineering.*

[15] Small Multi-Mission Satellite (SMMS) Data Retrieved: May 26, 2010 from the World Wide Web. MAMBO. [Online]. Available: http://smms.ee.ku.ac.th/index.php

[16] R. R. Bouckaert, "WEKA Manual," WAIKATO University, pp. 1-303, January 2010. Retrieved: May 26, 2010 from the World Wide, I Table Command Line [Online]. Available: www.cs.uu.nl/docs/vakken/dm/WekaManual.pdf

[17] J. R. Jensen, "Introductory Digital Image Processing--A Remote Sensing Perspective," *Prentice Hall, Inc, New Jersey*, pp. 197-256, 1996

Kitti Koonsanit received his M.S. degree in computer science from Kasetsart University in 2008. He is currently a Ph. D. candidate in computer science, Kasetsart University, Thailand. His fields of interest include clustering, image processing, image segmentation, band selection, multispectral image, and medical imaging.

Chuleerat Jaruskulchai received her D.Sc. degree in computer science from George Washington University, School of Engineering and Applied Science, USA in 1998. She is currently an Associate Professor and lecturer in the Department of Computer Science, Kasetsart University, Thailand. Her fields of interest and research areas include information retrieval, clustering, text classification, and statistic modeling.

Apisit Eiumnoh received his Ph.D. degree in Soil Genesis & Classification from North Carolina State, UK. He is currently an Associate Professor in National Center for Genetic Engineering and Biotechnology, Patumthani, Thailand.