Unsupervised Image Classification for Large Remotely-sensed Imagery using Region-growing Segmentation

Sang-Hoon Lee

Kyungwon University

Abstract – A multistage hierarchical clustering technique, which is an unsupervised technique, was suggested in this paper for classifying large remotely-sensed imagery. The multistage algorithm consists of two stages. The local segmentor of the first stage performs region-growing segmentation by employing the hierarchical clustering procedure of CN-chain with the restriction that pixels in a cluster must be spatially contiguous. This stage uses a sliding window strategy with boundary blocking to alleviate a computational problem in computer memory for an enormous data. The global segmentor of the second stage has not spatial constraints for merging to classify the segments resulting from the previous stage. The experimental results show that the new approach proposed in this study efficiently performs the segmentation for the images of very large size and an extensive number of bands

1. Introduction

Most approaches to the image classification require *a priori* class-dependent knowledge of parameterized models for the data. In many instances, however, the parameter values of the models are not known *a priori*, and the process of gathering training samples to estimate parameters is often infeasible or very expensive. In addition, the classification results much depend on the number of classes selected in the specific analyzed area, but it is very complicate to determine the class number, as known as "cluster validation," particularly for remotely sensed data. Therefore, it is necessary that classification procedures perform the unsupervised learning of the parameters including the number of classes and the image classification simultaneously.

Due to advances in sensor technology, it is now possible to acquire high-resolution data over large geographical area. The high-resolution or hyperspectral imagery possesses much detailed spatial information, but one of challenging problems in processing this large dimensional data is the computational complexity resulting from processing the vast amount of data volume. Especially, the unsupervised classification that makes use of hierarchical clustering (Anderberg, 1973) may require enormous processing time for large images. Lee (2001) used a multistage classification approach based on regional growing, which is computationally efficient for the unsupervised classification. The multistage algorithm includes two stages of segmentation. The first stage performs region-growing segmentation that confines merging to spatially adjacent clusters and then generates an image partition such that no union of any neighboring segments is uniform The "local" segmentor employs a hierarchical clustering procedure that merge the best "mutual closest neighbor (MCN)" pair satisfying a given clustering criterion. In the second stage, the image partition resulting from the regional segmentation is classified into a small number of distinct states by a sequential merging operation. To alleviate the memory problem and improve the computational performance of the algorithm, this approach uses a sliding window strategy of boundary blocking operation. The sliding window strategy is outlined in Fig. 1.

2. CN-chain Spatial Clustering

The computational efficiency of hierarchical clustering segmentation is mainly dependent on how to find the best pair to be merged. Let $I_n = \{1,2,\cdots,n\}$ be an index set of pixels of a sample image, $J_M = \{1,2,\cdots,M\}$ be an index set of regions associated with $\mathbf{G}_J = \{G_j \subseteq I_n \mid j \in J_M\}$ that is a partition of I_n, $\mathbf{R}_J = \{R_j \subseteq I_n \mid j \in J_M\}$ be a region neighborhood system such that R_j is the index set of neighborhood regions of region j. The closest neighbor of region j is defined as

$$\mathrm{CN}(j) = \arg\min_{k \in R_j} d(j,k) \tag{1}$$

where $d(j,k)$ is the dissimilarity measure between regions j and k, and R_j is the index set of regions considered to be merged with region j. The pair of regions is then defined as MCN iff $k =$

CN(j) and j = CN(k). If a cutting rule that

$$\mathrm{CR}(j,k) < \mathrm{CR}_{\max} \tag{2}$$

is given as a merging condition of two regions, for any arbitrary region r_0 satisfying $\mathrm{CR}(r_0, r_1) < \mathrm{CR}_{\max}$, a CN-chain is established as the sequence of regions

$$r_0, r_1 = \mathrm{CN}(r_0), \cdots, r_{h-1} = \mathrm{MCN}(r_h) \tag{3}$$

such that the last two region constitutes an MCN pair, i.e.

$$r_{h-1} = \mathrm{CN}(r_h) \text{ and } r_h = \mathrm{CN}(r_{h-1}).$$

The sequence of (3) also satisfies

$$\mathrm{CR}(r_{k-1}, r_k) < \mathrm{CR}_{\max} \text{ for } k = 1,2,\cdots,h.$$

The CN-chain algorithm is outlined in the following:

1. Initialize that every pixel is defined as a region, i.e., $J_M \leftarrow I_n$.

2. Construct a CN-chain by starting the region r_0 with the lowest index among J_M satisfying

$$\mathrm{CR}(r_0, \mathrm{CN}(r_0)) < \mathrm{CR}_{\max}.$$

   If there exists no available region, **STOP**.

3. Merge the last MCN pair, r_{h-1} and r_h by indexing the new region with the lower one of both region indices and update the partition by eliminating the region with the higher index, i.e.

$$r_{new} \leftarrow \min\{r_h, \mathrm{MCN}(r_h)\}$$
$$J_M \leftarrow J_M \setminus q \text{ where } q = \max\{r_h, \mathrm{MCN}(r_h)\},$$

   and update the regional parameters of the new region r_{new}.

4. For $h = 1$,
   if $\mathrm{CR}(r_0, \mathrm{CN}(r_0)) < \mathrm{CR}_{\max}$, reconstruct the CN-chain from the starting region r_0.
   else **GOTO** Step 2.

   For $h > 1$,
   if $\mathrm{CR}(r_{h-2}, \mathrm{CN}(r_{h-2})) < \mathrm{CR}_{\max}$, reconstruct the CN-chain from the starting region r_{h-2}.
   else **GOTO** Step 2.

5. **GOTO** Step 3.

One common objective in image segmentation involves minimizing the over-all intra-cluster sample variance. This results in the maximum likelihood solution for the case of unknown parameters in a Gaussian image field. Other statistical measures can also be employed to obtain well-posed solutions, but at a higher computational cost (Lee and Crawford, 2005). The advantages of intra-cluster sample variance are both simplicity and its ability to represent a basic important characteristic of clusters. In the local segmentation of spatial region-growing, this study assumes a simple variance structure with no correlation between bands for the image data processes. Under this assumption, the dissimilarity measure based on the intra-cluster sample variance is defined:

$$d(r,s) = \frac{n_r n_s \sum_{k=1}^{b} (\hat{\mu}_{rk} - \hat{\mu}_{sk})^2}{n_r + n_s} \tag{5}$$

where b is the number of bands, n_j and $\hat{\mu}_{jk}$ are the number of pixels and the kth band's average of G_j respectively.

It is difficult to define an appropriate measure of homogeneity to establish rules for cutting the hierarchical tree. The model fitting approaches using information criteria which measure the trade-off between the likelihood and the penalty for increasing the model's order have been applied for cluster validation in image analysis (Won and Derin, 1992). One of these approaches involves selecting the optimal state that maximizes Schwarz's information criterion (Schwarz, 1978):

$$\mathrm{SIC} = -2\log L(h) + K(h)\log n \tag{6}$$

where h is the number of distinct states, $L(h)$ and $K(h)$ are the maximum value of the likelihood function and the number of independent parameters estimated when using h states respectively. The SIC statistic was derived assuming a nonzero prior based on an asymptotic approximation to Bayes' loss. Although the Schwarz's approximation may fail for small samples because of its asymptotic nature, the use of SIC is generally appropriate for selecting the number of classes in image analysis. According to the SIC, merging of two regions can be considered if the decrease in the log likelihood is less than ($0.5\log n$) multiplied by the change in the number of parameters associated with a merged class.

3. Experiments and Conclusion

The proposed CN-Chain Classification (CNCC) was evaluated using simulation data generated by the Monte Carlo method. Table 1 shows performance of CNCC in CPU time for the simulation data of different sizes and band-numbers. The experiments with simulation data show that the use of CN-chain efficiently performs the segmentation for the images of very large size and an extensive number of bands.

Table 1. CPU Times of CNCC for different image sizes with 3 bands and different numbers of bands with 1K x 1K.

Size	CPU Time (sec)
2K x 2K	36.98
4K x 4K	166.03
10K x 10K	1211.02

Number of Bands	CPU Time (sec)
10	22.84
50	202.93
200	485.95

References

Anderberg, M. R, 1973. *Cluster Analysis for Application*, Academic Press, NY.

Lee, S-H., 2001. Unsupervised image classification using spatial region growing segmentation and hierarchical clustering, 17:57-70 (in Korean).

Lee, S-H. and M. M. Crawford, 2005. Unsupervised Multistage Image Classification Using Hierarchical Clustering with a Bayesian Similarity Measure, *IEEE Trans. image processing*, 14: 312-320.

Schwarz, G., 1978. Estimation of the Dimension of a Model, *Annal., Math., Statist.*, 6: 461-464.

Won, C. S. and H. Derin, 1992. Unsupervised segmentation of noisy and textured images using Markov random fields, *Comp. Vision, Graphics, Image Processing*, 54: 308-328.

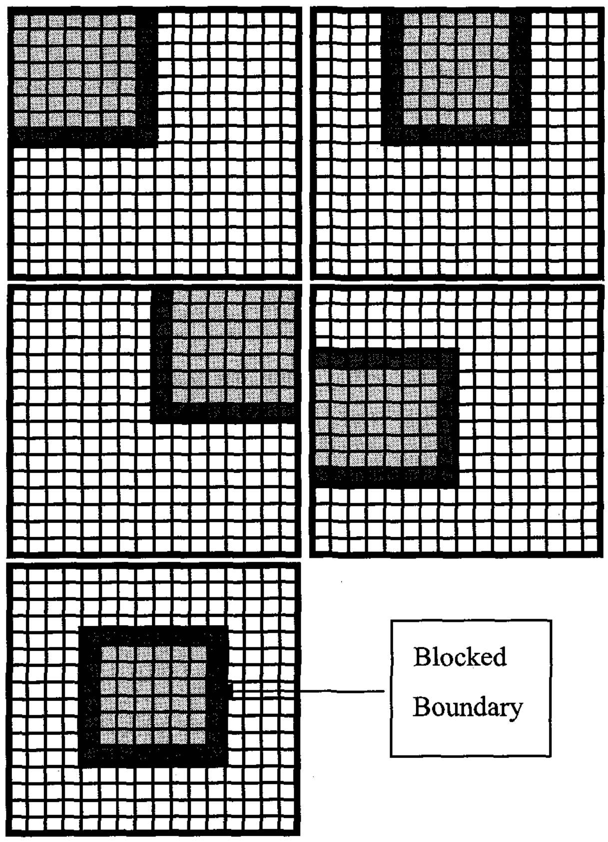

Fig. 1. Outline of sliding window strategy.