# An Improvement of KMeans Algorithm Using Wavelet Technique to Increase Speed of Clustering Remote Sensing Images

Trung Nguyen Tu[1*], Huy Ngo Hoang[1], Duc Dang Van[1], Thoa Vu Van[2], An Dang Duy[3]
[1] Institute of Information Technology, Vietnamese Academy of Science and Technology, Hanoi, Vietnam.
[2] Post & Telecommunications Institute of Technology, Hanoi, Vietnam.
[3] Institute of Industrial Machinery and Instruments Holding, Vietnam.

* Corresponding author. Tel.: 0936114331, 0966550565; email: trungnt.sremis@gmail.com
Manuscript submitted October 23, 2015; accepted March 26, 2016.
doi: 10.17706/ijcee.2016.8.2.177-184

**Abstract:** Remote sensing image clustering is the issue that is interested by remote sensing researchers. Remote sensing image can have multi bands and high space resolution. There are multi algorithms as K-Means, C-Means, Watersed... Wherein, KMeans is used and applied commonly to cluster remote sensing images about clustering quality and performance. However, when clustering large size remote sensing images, the converging speed of the algorithm is still slow. Moreover, this algorithm only points out clustered attribute of each pixel but do not show center near level of pixels in each cluster. This paper presents a technique which combines the algorithm Kmeans with Wavelet technique to execute remote sensing image with large and adding weights to adjust the center computing formula.

**Key words:** Clustering, remote sensing images, Kmeans, wavelet, weight.

## 1. Introduction

Remote sensing image processing in general and remote sensing image clustering in particular are the problems researched a long time ago and they are now concerned. The segmentation is a process which extracts the outline of the ground objects by defining homogenous regions [1]. The mission of remote image clustering function is that dealing with and dividing an initial image into different areas. At the moment, there are many different clustering methods such as Morphological methods, K-means, Fuzzy CMeans, … Most of the methods just use the intensity of each pixel to definite the areas, but they give a mish-mash of partitions, in detail, for multispectrum images with high resolution. Nowadays, some algorithms now include contextual information in the process to reduce the heterogeneity of the segmentations [1]. Wherein, some of them textural information extracted from the image is also used [1]. In [2], Chen and coworker proposed a new KMeans clustering algorithm using center displacement. In [3], Balaji and coworker presented a new clustering algorithm based on converting image from RGB color space to L*a*b color space and clustering in this space. In [4], the authors used Wavelets to reduce the noise of medical images.

If comparing with some clustering algorithms, KMeans algorithm combines the strengths: faster speed, center cluster control and effective clustering even with big images. Perhaps, this is reason why KMeans

have been used popularly in research and installed in sensing image processing softwares. However, when clustering the large size sensing images, the converging speed of algorithm is still slow. In [2], the authors proposed the CCEA center initing algorithm to increase the speed of fuzzy KMeans algorithm. In [5], [6], authors proposed the weighted KMeans algorithms. However, this algorithm just shows property of belong to a cluster for each pixel, but don't show clearly the center near level of pixels in each cluster.

In this research, we propose an improvement of K-Means clustering algorithm which combines between K-Means algorithm and Wavelet technique in effective center initalizing to increase the speed of clustering large images in general and remote sensing images in particular and supplement weight to adjust the formula for calculating the cluster center.

## 2. KMeans Clustering Algoritm

### 2.1. KMeans Algorithm

KMeans algorithm [7] includes 4 steps. Input: $n$ object and number of clusters $k$. Output: Clusters $C_i$ ($i$=1..$k$) so that the follow objective function $E$ reachs minimum: $E = \sum_{i=1}^{k} \sum_{x \in C_i} d^2(x, m_i)$ (1). Step 1: Initialization. Step 2: Assigning cluster center according to distance with (2). Step 3: Updating the cluster center with (3). Step 4: Repeating and testing condition to stop.

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} d^2(x, m_i) \quad (1)$$

$$d(x, C_S) = \min d(x, C_j), 1 \le j \le k \quad (2)$$

$$C_j = \frac{\sum_{x \in \text{cluster}(j)} x}{\text{count}(\text{cluster}(j))} \quad (3)$$

### 2.2. Analyzing Issues of KMeans Clustering Algorithm

Although KMeans algorithm effectively clusters remote sensing images, there are 2 issues:

- Issue 1: Enforcement speed for big images like remote sensing images is slow.

With KMeans algorithm [7], step 1, initializing randomly the cluster centers affects the speed of algorithm. Each time that algorithm executed, if the random choosed centers are good, suppose they are near to the positions of centers after converging, enforcement time left is not great. However, if the centers are not good, suppose they are very near to each other. At that moment, it will waste much time to determine the centers after converging because of many repeating times. This is the reason why researchers find the ways so that the cluster centers is initialized the best.

- Issue 2: Not clearly distinguishing the role of pixels in the same cluster with the cluster center.

In image clustering, an object is an pixel. According Table 1, with cluster garthering criterion (step 2) and updating the cluster centers (step 3), we can make the following comments. The first, the algorithm shows the relationship in the same cluster of objects (cluster garthering criterion) in each cluster (*). The second, the algorithm does not distinguish clearly the level of effect, near center of pixels in each cluster (cluster center updating) (**). We can see the 2nd comment (**) through illustration in Fig. 1. See in Fig. 2, we see that P1 and P2 will be brought together in cluster with C1 core because $d(P1,C1) < d(P1,C2)$ and $d(P2,C1) < d(P2,C2)$. Call d1 = $d(P1,C1)$ and $d2 = d(P2,C1)$. Suppose $d2 < d1$, it means that P2 is nearer to the center C1 than $P1$. Therefore, $P2$ has to have more affected than $P1$ in $C1$ cluster. However, the formula of updating cluster centers does not reflect this. In addition, if using the formulate of calculating the cluster center like KMeans, it can be affected by noises because the roles of pixels (may be noise or not) are the same.

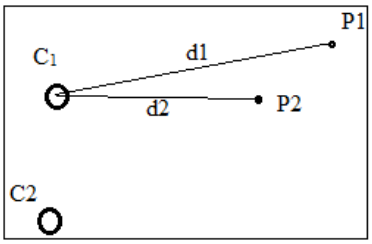



Fig. 1. The roles of pixels in cluster according to distance criterion.

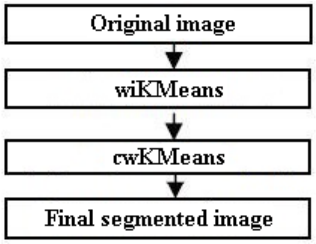



Fig. 2. Diagram of wwKMeans algorithm.

## 3. Improved KMeans Clustering Algorithm

### 3.1. Proposing Improved Algorithm

In this part, we propose the improved KMeans clustering algorithm for sensing images which we temporarily call wwKMeans (wavelet initialized and center weight KMeans). wwKMeans is combined from two algorithms: wiKMeans (wavelet initialized KMeans) and cwKMeans (center weight KMeans). The mission of wiKmeans is to make a good initializing center set to solve issue 1. cwKMeans use the formula of center calculating is adjusted to solve issue 2. The algorithm diagram is illustrated in Fig. 3.

From diagram in Fig. 2, the algorithm is enforced as follow. The original image is brought into the wiKMeans algorithm to make the initialized center set. This center set is used to be input with the original image for cwKMeans algorithm to make the final segmented image.

### 3.2. The Cluster Center Initializing Algorithm wiKMeans

To overcome the limitation (1), in this subsection, we propose the cluster center initializing solution which combines Wavelet and KMeans algorithm. Discrete Wavelet transformation is used to decrease the size of input image. Minima approximated image of Wavelet transformation is used to clustering to obtain a center set. The algorithm diagram is illustrated in Fig. 3.

**Step 1: Discrete Wavelet Transform**

*Wavelet Transform (WT)* is a mathematic tool which is usually used to present multi-resolution image. After transform, we obtain Wavelet coefficient set. WT can be presented similarly with Fourier transform.

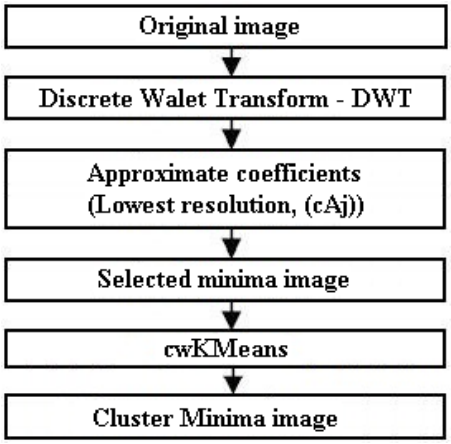



Fig. 3. Diagram of wiKmeans algorithm.

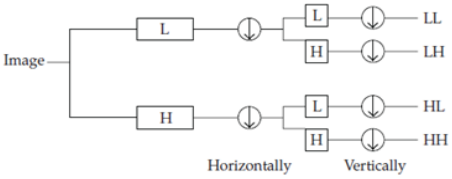



Fig. 4. Image transformation with Wavelet.

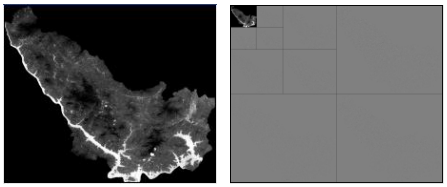


Fig. 5. Input image and the disaggegated result.

With digital signal like remote sensing images, Wavelet coefficient set can be collected by Discrete Wavelet Transform (*DWT*). For most of digital images, the content of low frequency is the most important, which keeps most the features of input image with size decrease by 4 times. After applying low pass filter

according to 2 directions (LL) we get the approximate image of the original image. Repeating the above process with (LL) to make coefficients in 2nd level. Fig. 4 illustrates DWT to image according to pyramid algorithm of Mallat [8].

Each time DWT is perfromed, the size of approximate image LL is 4-time smaller than the one of the previous. Therefore, suppose we disaggregate 3 levels for the input image, we get approximate image which has the size decreased 64 times. Fig. 5 illustrates input image and disaggregated result.

**Step 2: clustering minima image by K-means**

Clustering the minima approximated image with KMeans algorithm. After clustering, we have the follow cluster center set:

$$V_{\text{Init}} = \{v_k: \ 1 \le k \le c\} \tag{4}$$

*Comment:*

Because KMean algorithm in wiKMeans is not enforced directly on the original image, but on smaller image, the enforcement speed is faster and it is faster and faster if the level of disaggregation is high. Moreover, the input image of KMeans is the approximate image of the original image so almost the features of original image are kept, the main difference is size and noise rate. Therefore, center set gathered when KMeans converges in wiKMeans near to center set when coverging in the case which KMeans clusters directly on the input image.

## 3.3. Image Clustering Algorithm cwKMeans with New Cluster Center Updating Formula

In order to overcome the limitation (2), in this subsection, we propose the formula of calculating new cluster center as (5). Wherein, $x$ is object (pixel) belongs to cluster $j$, $C_j$ is cluster center $j$, $d(x, c_j)$ is distance between object (pixel) $x$ and center $C_j$. In this study, we use Euclide distance.

$$C_j = \frac{\sum_{x \in cluster(j)} x \times \frac{1}{d(x, C_j)}}{\sum_{x \in cluster(j)} \frac{1}{d(x, C_j)}} \tag{5}$$

$$d_\varepsilon(x, c_j) = \begin{cases} d(x, c_j) & \text{if } d(x, c_j) \ge \varepsilon \\ \varepsilon & \text{if } d(x, c_j) < \varepsilon \end{cases} \tag{6}$$

$$C_j = \frac{\sum_{x \in \text{cluster}(j)} x \times \frac{1}{d_\varepsilon(x, c_j)}}{\sum_{x \in \text{cluster}(j)} \frac{1}{d_\varepsilon(x, c_j)}} \tag{7}$$

$$C_j = \frac{\sum_{x \in \text{cluster}(j)} x \times \frac{1}{d_\varepsilon(x, c_j)}}{\sum_{x \in \text{cluster}(j)} \frac{1}{d_\varepsilon(x, c_j)}} \tag{8}$$

In this case, if object $x$ matches (same coordinates) to center $C_j$, then $d(x, c_j) = 0$. Because this distance plays role of denominator, it cannot have value zero. Therefore, we propose $d_\varepsilon(x, c_j)$. Wherein, $\varepsilon$ is a threshold. $d_\varepsilon(x, c_j)$ is calculated as (6). In this study, we experiment with $\varepsilon = 0.000001$. Meanwhile, (9) is rewritten as (11):

From proposing the formula (7) and the result comment of wiKMeans algorithm (subsection 3.2, we also

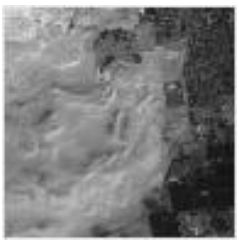
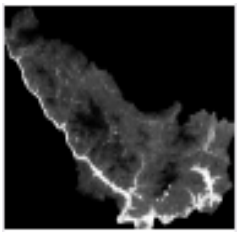
propose the Kmeans based improved clustering algorithm, supplementing weight, which is called cwKMeans (center weight KMeans), includes 4 steps as KMeans but step 3 using formula (8).

## 4. Experiments

We test the proposed algorithm wwKMeans. CCEA center initializing algorithm [2] helps to increase the speed of KMeans. Wherein, we also compared the enforcement speed of wwKmeans with CCEA_Kmeans algorithm. We disaggregate Wavelet in 3 levels with kernel function Biorthogonal. Dataset used for experiments includes 3 types: Landsat ETM+ and SPOT (high resolution) images taken in Hoa Binh and Son La, Quickbird images are downloaded from model data on website: http://opticks.org. Because of the limited scope of the paper, we present experiments with different two input images.

In 1st experiment, the original image is Quickbird with its size of 2056 ×2065 (pixels). In 2nd experiment, the original image is LANDSAT with its size of 1596 × 1333. Table 2 illustrates images on Green, Indigo channel of image samples in experiments 1, 2.

Table 1. Input Images in 1st, 2nd Experiments

| Experiment 1 | Experiment 2 |
|---|---|
|  |  |

### 4.1. Expriment 1

Table 2 illustrates the clustering result of KMeans, CCEA_KMeans and wwKMeans algorithm on Blue channel in the case of five clusters.

Table 2. Image of Result on Blue Channel Segmented by Kmeans, CCEA_kmeans and wwKmeans with 5 Clusters

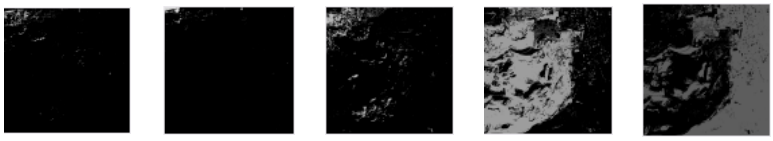
| Cluster | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| KMeans |  | | | | |
| CCEA_KMeans |  | | | | |
| wwKMeans |  | | | | |

Table 3 shows clustering time of KMeans, CCEA_KMeans and wiKMeans algorithms with the number of clusters 5, 7, 9, 12 and 16 in turn.

Table 3. Compare Clustering Time (ms) of Kmeans, CCEA_Kmeans and wwKmeans Algorithms with Different Number of Clusters

| No Cluster | KMeans | CCEA_KMeans | wwKMeans |
|---|---|---|---|
| 5 | 2339797 | 213828 | 111047 |
| 7 | 2309922 | 219203 | 148141 |
| 9 | 7092406 | 369016 | 257890 |
| 12 | 13765729 | 463422 | 344860 |
| 16 | 18985531 | 611750 | 518343 |

Tables 4, 5 show disaggregated measure of clusters through distance of centers of cwKMeans and KMeans algorithms in the case of five clusters.

Table 4. Distance between Centers Created from cwKmeans

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 68.08 | 111.562 | 180.03 | 237.51 |
| 2 | | 0 | 51.94 | 120.45 | 173.86 |
| 3 | | | 0 | 68.53 | 132.99 |
| 4 | | | | 0 | 73.5 |
| 5 | | | | | 0 |

Table 5. Distance between Centers Created from Kmeans

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 67.05 | 111.557 | 179.38 | 240.52 |
| 2 | | 0 | 45.04 | 113.39 | 172.52 |
| 3 | | | 0 | 68.37 | 127.47 |
| 4 | | | | 0 | 68.18 |
| 5 | | | | | 0 |

## 4.2. Expriment 2

Table 6 illustrates the clustering result of KMeans, CCEA_KMeans and wwKMeans algorithm on Indigo channel in the case of five clusters.

Table 6. Image of Result on Blue Channel Segmented by Kmeans, CCEA_kmeans and wwKmeans with 5 Clusters

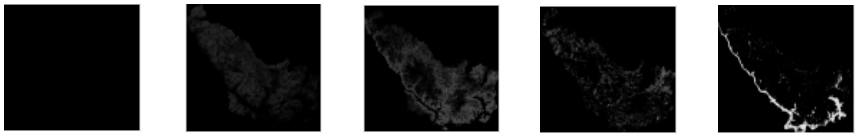
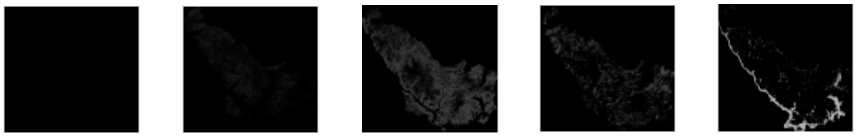
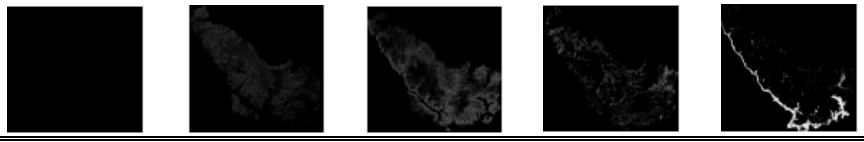
| Cluster | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| KMeans |  | | | | |
| CCEA_KMeans |  | | | | |
| wwKMeans |  | | | | |

Table 7 shows clustering time of KMeans, CCEA_KMeans and wwKMeans algorithms with the number of clusters 5, 9, 13, 18 and 21 in turn.

Table 7. Compare Clustering Time (ms) of Kmeans, CCEA_Kmeans and wwKmeans Algorithms with Different Number of Clusters

| No Cluster | KMeans | CCEA_KMeans | wwKMeans |
|---|---|---|---|
| 5 | 329453 | 66953 | 39656 |
| 9 | 2602187 | 141297 | 108172 |
| 13 | 2724187 | 181437 | 135291 |
| 18 | 8146656 | 303516 | 250328 |
| 21 | 9046766 | 340000 | 306703 |

Tables 8, 9 show disaggregated measure of clusters through distance of centers of cwKMeans and KMeans algorithms in the case of five clusters.

Table 8. Distance Between Centers Created from cwKmeans

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 109.09 | 198.87 | 116.79 | 368.63 |
| 2 | | 0 | 93.73 | 38.42 | 266.01 |
| 3 | | | 0 | 89.94 | 173.57 |
| 4 | | | | 0 | 252.94 |
| 5 | | | | | 0 |

Table 9. Distance Between Centers Created from Kmeans

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 107.68 | 196.19 | 115.72 | 366.51 |
| 2 | | 0 | 95.46 | 38.74 | 265.52 |
| 3 | | | 0 | 90.3 | 171.45 |
| 4 | | | | 0 | 249.39 |
| 5 | | | | | 0 |

*Comment:*

First, from Table 3 in 1st experiment, Table 7 in 2nd experiment, we can see that enforcement time of wwKMeans is smaller than CCEA_KMeans and smaller than very much KMeans. Second, from Tables 4, 5 in 1st experiment, Tables 8, 9 in 2nd experiment, disaggregated measure of clusters of cwKMeans is better than disaggregated measure of clusters of cwKMeans of KMeans.

## 5. Conclusion

In this research, we proposed wwKmeans algorithm, including two small algorithms: wiKmeans and cwKMeans. wiKmeans to initialize center set well and help increase the speed of clustering remote sensing images. cwKMeans to cluster with new fomula of calculating cluster centers to distinguish the effect level of pixels in the same cluster and decrease noises. wwKmeans uses wiKmeans to get the good initialized center set. Next, using cwKmeans to cluster the original image with the above center set. The results of experiments show 2 things. First, enforcement time of wwKMeans is smaller than CCEA_KMeans and smaller than very much KMeans. Second, disaggregated measure of clusters of cwKMeans is better than disaggregated measure of clusters of cwKMeans of KMeans.

In the next work, the group will test with continuing wavelet transformation with different kernel function.

## References


[1] Cuadra, M. B., & Thiran, J.-P. (2004). Satellite image segmentation and classification. Diploma Project.

[2] Chang, C.-T., Lai, J. C., & Derjeng, M. (2011). A fuzzy k-means clustering algorithm using cluster center displacement. *Journal of Information Science and Engineering*, *27*, 995-1009.

[3] Balaji, T., & Sumathi, M. (August 2013). Relational features of remote sensing image classification using effective k-means clustering. *International Journal of Advancements in Research & Technology*, *2(8)*, 103-107.

[4] Dalmiya, S., Dasgupta, A., & Datta, S. K. (2012). Application of wavelet based k-means algorithm in mammogram segmentation. *International Journal of Computer Application*, *52(15)*.

[5] Kerdprasop, K., Kerdprasop, N., & Sattayatham, P. (2005). *Weighted K-Means for Density-Biased Clustering*. Springer-Verlag Berlin Heidelberg, 488-497.

[6] Ackerman, M., Shai, B.-D., Branzei, S., & Loker, D. *Weighted Clustering*. From http://cs.au.dk/~simina/weighted.pdf

[7] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*: *Vol. 1* (pp. 281-297).

[8] Mallat, S. G. (1989). A theory for multi resolution signal decomposition, the wavelet representation.

*IEEE Transactions on Pattern Analysis and machine Intelligence*, *11(7)*, 674-693.

**Trung Nguyen Tu** was born in 1985 at Hanoi. Trung graduated from Hanoi Pedagogical University in 2007 and completed his master degree from University of Engineering and Technology in 2011. Trung is a doctoral candidate at the Post & Telecommunications Institute of Technology. His research interests are image processing, speech processing, information systems, and embedded systems.

**Huy Ngo Hoang** was born in 1969 at Hanoi. Huy graduated from Hanoi Pedagogical University in 1990. He is a doctoral candidate at the Institute of Information Technology, Vietnamese Academy of Science and Technology. He also works at the same institute. His research interests are image processing, speech processing, information systems, embedded systems, and automatic control.

**Duc Dang Van** was born in 1951. He was a doctoral candidate at the Institute of Information Technology, Vietnamese Academy of Science and Technology and finished in 1996. He became an associate professor in 2002. He works at the Institute of Information Technology, Vietnamese Academy of Science and Technology. His research interests are image processing, speech processing, GIS, remote sensing image processing, multimedia, and software engineering.

**Thoa Vu Van** was born in 1955 at Ninh Binh. He graduated from Vinh Pedagogical University in 1975. He received a PhD degree in 1990. He works at the Post & Telecommunications Institute of Technology. His research interests are algorithm theory, optimization, geographic information system, and network.

**An Dang Duy** was born in 1974 at Thai Nguyen. He graduated from Hanoi University Science and technology in 2001. He received his master degree from Thai Nguyen University in 2014. He works at the Institute of Industrial machinery and Instruments Holding. His research interests are image processing and speech processing.