UNSUPERVISED CLASSIFICATION OF SATELLITE IMAGES USING KHM ALGORITHM AND CLUSTER VALIDITY INDEX

Habib Mahi[1], Nezha FARHI[2] and Kaouther LABED[2]

1. Earth Observation Department, Centre of Space Techniques, Arzew, Algeria; hmahi@cts.asal.dz, nfarhi@cts.asal.dz
2. Faculty of Mathematics and Computer Science, Mohamed Boudiaf University – USTOMB, Oran, Algeria; kaouther.labed@univ-usto.dz

ABSTRACT

In this paper, we present a process which intended to detect the optimal number of clusters in multispectral remotely sensed images. The proposed process is based on the combination of both the K-Harmonic means and cluster validity index with an angle based method. The experimental results conducted on both synthetic data sets and real data sets confirm the effectiveness of the proposed methodology. On the other hand, the comparison between the well-known K-means algorithm and the K-Harmonic means shows the superiority of this latter.

INTRODUCTION

In remote sensing applications, the unsupervised classification, also called clustering is an important task, which aims to partition the image into homogeneous clusters (1). In general, each cluster corresponds to a land cover type. The most commonly used algorithms in remote sensing are the K-Means (KM) (2) and ISODATA (Iterative Self-Organizing Data Analysis Technique) (3). Their popularity is mainly due to their simplicity and scalability, indeed, the user must specify only the number of classes in the image. However, it is difficult to have a priori information about the number of clusters in satellite images; so, it is necessary to determine this value automatically. On other hand, the KM algorithm and similarly the ISODATA algorithm work best for images with clusters which are spherical and that have the same variance. This is often not true for remotely sensed data, which are elongated with a larger variability, such as forest for example (4).

In this paper, we propose a new clustering method based on the junction of K-harmonic means (KHM) clustering algorithm (5), cluster validity indices (6) and an angle based method (7) in order to classify satellite images. The choice of the KHM algorithm is motivated by its insensitivity to the initialization of the centers unlike KM and ISODATA. In addition, a cluster validity index is introduced to determine the optimal number of clusters in the data studied. Five cluster validity indices were compared in this work namely, DB index (8), DB^* index (9), XB index (7), BIC index (10) and WB-index (11) and one of them is selected.

The adopted methodology consists to varying the number of clusters K from K^{min} to K^{max}, and then we compute the selected cluster validity index for each K for the result obtained using the KHM algorithm. The clustered image corresponding to the minimum value of the selected cluster validity index combined with the angle based method is presented as a best classification. The Experimental results and comparison with K-means (KM) algorithms confirm the effectiveness of the proposed methodology.

METHODS

This section presents an overview of clustering algorithm employed in this paper, namely K-Harmonic Means and introduces two clustering validity indices.

K-Harmonic Means Algorithm

K-Harmonic means clustering is an improved version of the K-Means that was proposed by Zhang in 1999 and 2000 (5). The KHM method being less sensitive to the initialization procedure than the KM algorithm. The KHM performance function is defined as:

$$KHM = \sum_{i=1}^{N} \frac{K}{\sum_{j=1}^{K} \frac{1}{\|x_i - c_j\|^q}} \quad \textbf{(1)}$$

New centers clusters are calculated as following (12):

$$c_k = \frac{\sum_{i=1}^{N} \frac{1}{\left[\sum_{l=1}^{K} \frac{\|x_i - c_j\|^q}{\|x_i - c_l\|^q}\right]^2} x_i}{\sum_{i=1}^{N} \frac{1}{\left[\sum_{l=1}^{K} \frac{\|x_i - c_j\|^q}{\|x_i - c_l\|^q}\right]^2}} \quad \textbf{(2)}$$

Validity indices

Validity indices are measures that are used to evaluate and assess the results of a clustering algorithm. In the following, we describe only two clustering validity indices among the five used in this work, namely Bayesian Information Criterion (BIC) and Cylindrical distance based Davies-Bouldin (DB*). More details of Davies-Bouldin (DB), Xie-Benie (XB) and WB index (WB) can be found in (13).

Bayesian Information Criterion (BIC↑) (7). Also known as the Schwarz Criterion, the BIC was introduced as a competitor to the Akaike Information Criterion. It is based in part on increasing a likelihood function and is formulated as follows:

$$\mathrm{BIC} = \sum_{i=1}^{M} \left(n_i \log \frac{n_i}{N} - \frac{n_i \times D}{2} \log(2\pi) - \frac{n_i}{2} \log \sum_{i} - \frac{n_i - M}{2}\right) - \frac{1}{2} M \log N \quad (3)$$

Where,

$$\sum_i = \frac{1}{N - M} \sum_{j=1}^{n_i} ||x_j - c_i||^2 \quad (4)$$

Davies-Bouldin based on Cylindrical distance index (DB ↓) (9).* This variation of the DB was proposed by JCR Thomas bringing a new measure called the cylindrical distance (9). The index tries to overcome the limitations of the Euclidean distance and is defined as follows:

$$\mathrm{DB}^* = \frac{1}{k} \sum_{i=1}^{k} max_{j \neq i} \left\{\frac{S_i + S_j}{\Theta_{(r,c_i,c_j)}}\right\} \quad (5)$$

Where,

$$\Theta_{(r,c_i,c_j)} = \frac{D_{i,j}}{|C| + 1} \quad (6)$$

Angle based method

When detecting the optimal number of clusters in a predefine range of an index values, we are often faced with local minimum or maximum problem depending on the index nature. Although

studies like combining the advantageous aspect of K-Harmonic means algorithm and Cluster validity indices can be used to solve optimization problems by choosing the first significant value, strong evidences in (7) proves that a good knee point detection method gives more accurate results if the right threshold is defined.

$$DiffFun(m) = F(m-1) + F(m+1) - 2 * F(m) \quad (7)$$

Where *DiffFun* represents the successive differences in the index function values F(m). In each curve, there are at least two obvious peaks (differences). In order to select the optimal local knee (peak) corresponding to the correct number of clusters, the angle propriety of the curve is used with the following formula (7):

$$Angle = atan\left(\frac{1}{|F(m) - F(m-1)|}\right) + atan\left(\frac{1}{|F(m+1) - F(m)|}\right) \quad (8)$$

In order to choose the best clustering validity index, the following procedure is performed:

1: **for** $k = k_{min}\ to\ k_{max}$ **do**

2: **for** $i = 1\ to\ 5$ **do**

3: Run the K-Harmonic Means Algorithm with k centers

4: Compute the value of CVI_i

5: **end for**

6: Select the best validity index in $CVI_{i \in \{1..5\}}$ using the angle based method.

7: **end for**

EXPERIMENTAL RESULTS AND DISCUSSION

In this section, series of tests are conducted in order to ensure the validity and effectiveness of the proposed method. All the experiments results have been obtained using the MATLAB software package.

Comparison between the four cluster validity indices

In order to select the best clustering validity index, we compared the five clustering validity indices using two different clustering algorithms, the well-known K-means algorithm and the K-Harmonic Means algorithm. We also employed during our evaluation, four 2D synthetic datasets, with the same number of objects and clusters (5000 objects, 15 clusters) and with different degree of overlapping, as depicted in Figure 1. These datasets are extracted from UCI Repository http://cs.uef.fi/sipu/datasets.

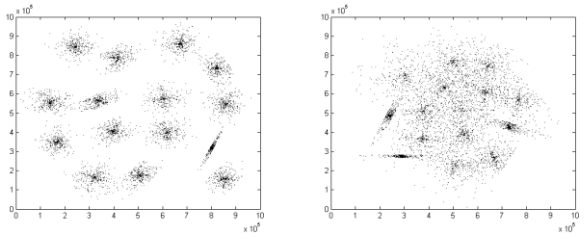

Figure 1: Synthetic data S1 and S4

Table 1. Comparison among the five CVI's for K-Harmonic Means using S1 dataset.

KHM	Without ABM	DB	XB	WB	DB*	BIC
		5	4	14	2	14
	With ABM	ADB	AXB	AWB	ADB*	ABIC
		14	14	14	16	16
K-Means	Without ABM	DB	XB	WB	DB*	BIC
		7	7	15	2	15
	With ABM	ADB	AXB	AWB	ADB*	ABIC
		15	15	15	15	15

Table 1 illustrates the efficiency of the angle-based method in order to find the best number of clusters. Usually, the used method selects the first significant minimum value. However, the results mentioned above show that the indices are very fluctuant making the returned results inaccurate even knowing that the clusters in the S1 dataset are well separated. We notice also, that the BIC returns the best value than the others with or without using the ABM. Regarding the used algorithms, they delivered approximately the same number of clusters with a little bit advantage of the K-means algorithm.

Table 2. Comparison among the five CVI's for K-Harmonic Means using S4 dataset.

KHM	Without ABM	DB	XB	WB	DB*	BIC
		5	5	15	3	3
	With ABM	ADB	AXB	AWB	ADB*	ABIC
		16	15	15	19	15
K-Means	Without ABM	DB	XB	WB	DB*	BIC
		4	5	15	15	5
	With ABM	ADB	AXB	AWB	ADB*	ABIC
		18	15	4	9	15

Table 2 shows the results for the highly overlapped dataset S4. The difference in the data distribution makes the CVI's values more fluctuant except for the WB. In this case, the first minimum value is not relevant of the correct number of clusters, making the use of the angle-based method necessary in order to approximate the right solution. As for the comparison between the five CVIs combined with the angle based method and the KHM algorithm, it is noticeable that the results are very close to the correct number of clusters in most cases. Unlike the combination of the method with the K-means that tends to return an incorrect number of clusters.

At the end, the combination of KHM algorithm, the angle properties and the CVIs is a very effective way to deal with local minima or maxima problems among a large range of datasets. Even considering some indices like the WB returns good results, the use of the angle-based method still provides a worthy amelioration on many indices such as the DB*. With regard to the previous ascertainment, we decided to choose the BIC index in order to apply our algorithm on remotely sensed datasets, knowing that all the indices give approximately the same number of clusters.

Experiment on Remotely Sensed Data

Besides the synthetic datasets, we used in the second experiment, three sub-scenes acquired by different sensors. The key characteristics of remotely sensed data used in this section are reported in Table 3.

Table 3. Key characteristics of remotely sensed datasets.

	Size	Resolution	satellite	Area

Sub scene 1	400 × 400	20m	Spot	Oran
Sub scene 2	500 × 500	10m	Alsat-2A	Tlemcen
Sub scene 3	600 × 800	30m	Landsat	Arzew

The clustering results of the three remotely sensed data by the proposed method are shown in Figure 2.d and Figure 2.f with seven clusters and Figure 2.e with four clusters, respectively. The obtained results appear generally satisfying according to the visual comparison with the corresponding original images. However, we notice confusion between water and shadow pixel, especially in case of the third image.

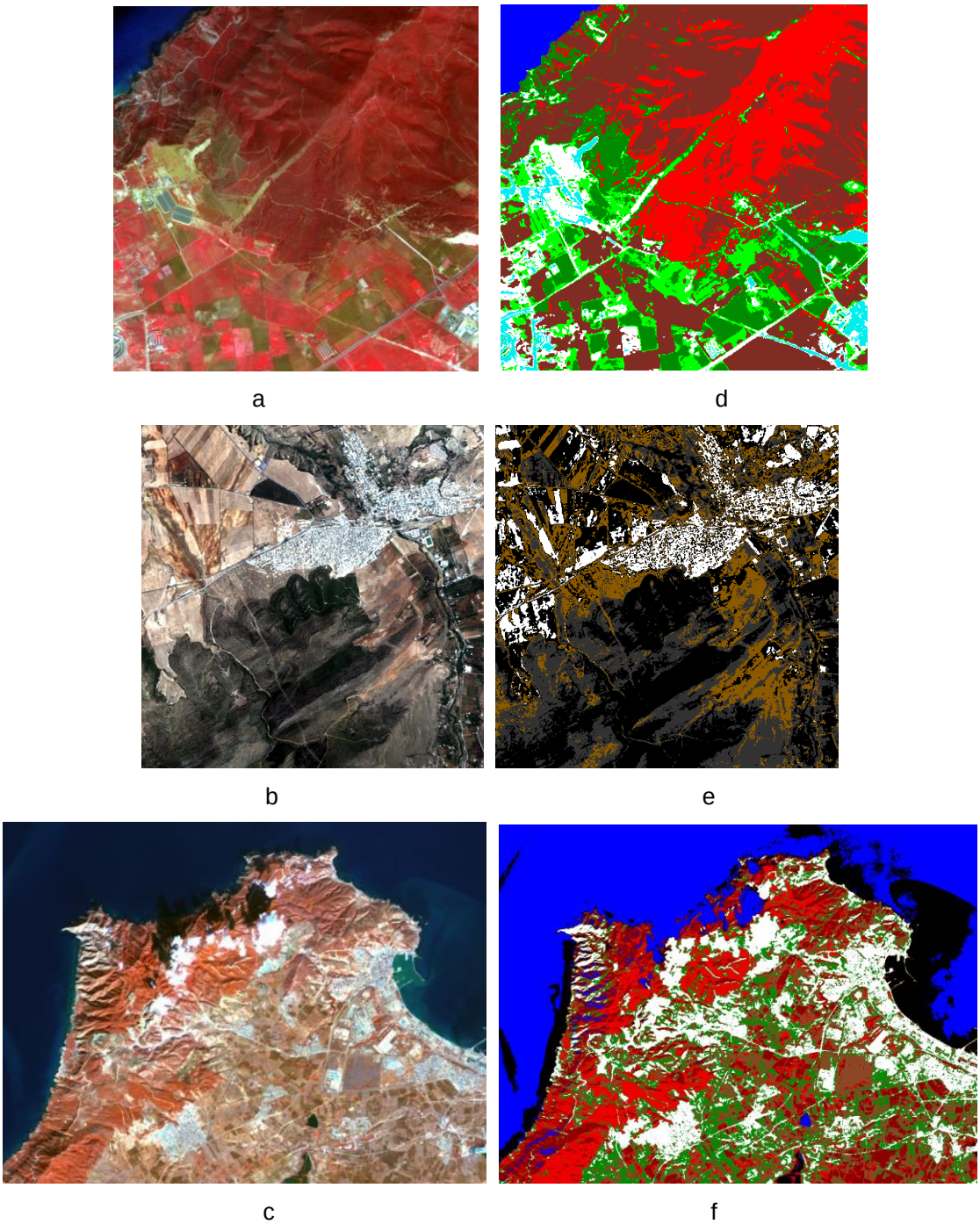

Figure 2: Clustering using the GKHM on remote sensed data sets

CONCLUSIONS

Combining an unsupervised classification method with cluster validity indices is a popular approach for determining the optimal number of clusters. In this paper, we proposed a combination of the KHM clustering algorithm, the cluster validity indices and an angle based method. The properties of each method are used and combined in order to improve the results by returning the most accurate number of clusters possible.

The experimental results section proves the efficiency of the proposed process against using a simple selection method by choosing the first significant minimum value. Other improvements could be done by testing the method on large datasets including high-dimensional datasets and shape sets.

A further research will involve the combination of both clusters validity indices and the angle based method with the GKHM [18] and finally the use of the ensemble clustering technique.

REFERENCES

1 G. Gan, C. Ma, and J. Wu, 2007. Data Clustering: Theory, Algorithms, and Applications. ASA-SIAM Serieson Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alex-andria, VA.

2 J. McQueen, 1967. Some methods for classification and analysis of multivariate observations. In Proc.5th Berkeley Symp Mathematics, statistics and probability, 281-296 pp.

3 G. Ball and D. Hall, 1965. ISODATA: A novel method of data analysis and pattern classification. In Technical report, Stanford Research Institute, Menlo Park, CA,USA.

4 GitanjaliS. Korgaonkar, R.R. Sedamkar and KiranBhandari, 2012. Hyperspectral Image Classification on Decision level fusion. IJCA Proceedings on International Conference and workshop on Emerging Trends in Technology. Vol. 7, 1-9 pp.

5 B. Zhang, 2000. Generalized K-Harmonic Means Boosting in Unsupervised Learning. Technical Reports, Hewllet Laborotories, HPL-2000-137.

6 M. K. Pakhira, S. Bandyopadhyay and U. Maulik, 2005. A Study of Some Fuzzy Cluster Validity Indices, Genetic Clustering and Application to Pixel Classification. Fuzzy Sets and Systems, vol. 155, 191-214 pp.

7 J.B.Talon S. Bourennane W.Philips D.Popescu P.Scheunders, 2008. Advanced Concepts for Intelligent Vision Systems. 10th International Conference, ACIVS 2008, Juan-les-Pins, France.

8 D. Davies and D. Bouldin, 1979. A cluster separation measure. IEEE PAMI, Vol. 1, no. 2, 224–227 pp.

9 J.C.R. Thomas, 2013. New Version of Davies-Bouldin Index for Clustering Validation Based on Cylindrical Distance. V Chilean Workshop on Pattern Recognition, Temuco, Chile.

10 X.L. Xie and A. Beni, 1991. Validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell. Vol.3, 841–846 pp.

11 Q. Zhao and P. Fränti, 2014. WB-index: a sum-of-squares based index for cluster validity. Knowledge and Data Engineering, Vol.92, 77-89 pp.

12 K. Thangavel and K. Karthikeyani Visalakshi, 2009. Ensemble based Distributed K- Harmonic Means Clustering. International Journal of Recent Trends in Engineering, Vol. 2, No. 1, 125-129 pp.

13 H. Mahi, N. Fargi, K. Labed, 2015. Remotely Sensed Data Clustering using K-Harmonic Means Algorithm and Cluster Validity Index. In : Computer Science and Its Applications - 5th IFIP TC 5 International Conference, CIIA 2015, Saida, Algeria, Vol 456, pp. 105-116.