UNSUPERVISED CLASSIFICATION OF SATELLITE IMAGES USING K-HARMONIC MEANS ALGORITHM AND CLUSTER VALIDITY INDEX

Habib Mahi[1], Nezha Farhi[1], and Kaouther Labed[2]

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

ABSTRACT

In this paper, we are presenting a process, which is 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 the latter.

KEYWORDS

Clustering, KHM, cluster validity indices, remotely sensed data, K-means, FCM.

INTRODUCTION

In remote sensing applications, the unsupervised classification, also called clustering, is an important task aiming to partition the image into homogeneous clusters (1,2). In general, each cluster corresponds to a land cover type. The most commonly used algorithms in remote sensing are the K-Means (KM) (3) and ISODATA (Iterative Self-Organizing Data Analysis Technique) (4). 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 (5). On the other hand, the KM algorithm and similarly the ISODATA algorithm work best for images with clusters which are spherical and have the same variance. This is often not true for remotely sensed data, where some clusters appear elongated in the feature space and different classes have different variability, e.g., forests tend to have larger variability than water (6).

In this paper, we propose a new clustering method based on the junction of K-harmonic means (KHM) clustering algorithm (7), cluster validity indices (8) and an angle-based method (9) in order to classify satellite images. The choice of the KHM algorithm is motivated by its insensitivity to the initialization of the centres unlike KM and ISODATA. In addition, a cluster validity index (CVI) is introduced to determine the optimal number of clusters in the data studied. Validity indices are measures that are used to evaluate and assess the results of a clustering algorithm. Five cluster validity indices were compared in this work, namely Davies Bouldin index (DB) (10), Cylindrical distance based Davies Bouldin index (DB*) (11), Xie Beni index (XB) (12), Bayesian Information Criterion (BIC) (5), and the sum of squares index (WB) (13) and one of them is selected.

METHODS

This section presents an overview of the clustering algorithm applied in this paper, namely K-Harmonic Means and introduces two clustering validity indices such as the BIC index and the DB* index. We notice that the adopted methodology is based on varying the number of clusters K from K_{min} to K_{max}, and then we compute the selected CVI for each K for the result obtained using the

DOI: 10.12760/01-2016-1-02

KHM algorithm. The clustered image corresponding to the minimum value of the selected CVI combined with the angle-based method is presented as the best classification.

The K-Harmonic Means Algorithm

The K-Harmonic means clustering algorithm is an improved version of the K-Means that was proposed by Zhang in 1999 and 2000 (7) and modified by Hammerly and Elkan in 2002 (14). The KHM method is less sensitive to the initialization procedure than the KM. The insensitivity to initialization is attributed to a dynamic weighting function, which increases the importance of the data points that are far from any centres in the next iteration (7). The KHM algorithm is given by:

Step 1: Acquire K initial centres c_j $(j = 1...K)$ among N data points and initiate $KHM^* = 0$

Step 2: Compute the value of the $KHM(X)$ performance function defined as:

$$KHM(X) = \sum_{i=1}^{N} \left(K \Big/ \sum_{j=1}^{K} \frac{1}{\left\| (x_i - c_j) \right\|^q} \right) \tag{1}$$

where: x_i is denotes an object in the input data set, q is a parameter and let $q \geq 2$

Step 3: Compute T_{ij} $(i = 1...N, j = 1...K)$ elements according to the following equation:

$$T_{ij} = \frac{\left\| (x_i - c_j) \right\|^{-q-2}}{\sum_{j=1}^{K} \left\| (x_i - c_j) \right\|^{-q-2}} \tag{2}$$

Step 4: Obtain the weight L_i of each data point given by:

$$L_i = \frac{\sum_{j=1}^{K} \left\| (x_i - c_j) \right\|^{-q-2}}{\left(\sum_{j=1}^{K} \left\| (x_i - c_j) \right\|^{-q} \right)^2} \tag{3}$$

Step 5: Update each cluster centres as following (15,16):

$$c_j = \frac{\sum_{i=1}^{N} T_{ij} L_i x_i}{\sum_{i=1}^{N} T_{ij} L_i} \tag{4}$$

Step 6: If $\left| KHM^* - KHM \right| > \varepsilon$, then $KHM^* = KHM$ and return to Step 2; otherwise go to Step 7

Step 7: Assign each data point x_i to the closest cluster c_j as follows:

$$j = \arg \max_{j=1...K} T_{ij} \tag{5}$$

Validity indices

In the following, we describe only two CVIs among the five used in this work, namely BIC and DB*. More details of DB, XB and WB index can be found in (17).

Bayesian Information Criterion (*BIC*↑) (5)

Also known as the Schwarz Criterion, the *BIC* index is similar to the Akaike Information Criterion (18). It is based in part on increasing the likelihood by adding more explaining variables and is formulated for clustering as follows:

$$BIC = \sum_{i=1}^{K} \left(n_i \log \frac{n_i}{N} - \frac{n_i \cdot d}{2} \log(2\pi) - \frac{n_i}{2} \log \Sigma_i - \frac{n_i - K}{2} \right) - \frac{1}{2} K \log N \tag{6}$$

Where, K represents the clusters
N is the size of the data set
n_i is the size of each cluster c_i
d is the dimension of the data sets.

Σ_i is the maximum likelihood estimated for the variance of the ith cluster as follows :

$$\Sigma_i = \frac{1}{N-K}\sum_{j=1}^{n_i}\left\|x_j - c_i\right\|^2 \tag{7}$$

where, x_j denotes an object in the input Data set and c_i represents the centroid of the ith cluster. High values of the BIC are strong evidences for good clustering results, so the index needs to be maximized in order to achieve best clustering.

Davies-Bouldin based on Cylindrical distance index (DB^* ↓) (11)

This variation of the *DB* was proposed by JCR Thomas introducing a new measure called the cylindrical distance (11). The index tries to overcome the limitations of the Euclidean distance and is defined as follows:

$$DB^* = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left\{\frac{S_i + S_j}{\Theta_{(r,c_j,c_i)}}\right\} \tag{8}$$

where S_i denotes the average distance between each point in the i^{th} cluster and the centroid of the i^{th} cluster, S_j denotes the average distance between each point in the i^{th} cluster and the centroid of the j^{th} cluster, and $\Theta_{(r,c_j,c_i)}$ denotes the cylindrical distance given by the following equation:

$$\Theta_{(r,c_j,c_i)} = \frac{D_{i,j}}{|C|+1} \tag{9}$$

where $D_{i,j}$ represents the Euclidean distance between the centroids of the i^{th} and j^{th} clusters, C denotes the subset of data points belonging to the region r and $|C|$ corresponds to its cardinality. Low values of the DB^* indicate good clustering results so the index should be minimized.

Angle-based method

When detecting the optimal number of clusters in a predefined range of index values, we are often faced with local minimum or maximum problems depending on the index nature. Although studies 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 (9) prove that a good knee point (peak) detection method gives more accurate results if the right threshold (δ) is defined. This method allows finding CVI tendencies by detecting the highest change in the index curve values. Different knee points summarize these changes. A threshold (δ) is defined in order to keep only significant peaks.

$$DiffFun(m) = F(m-1) + F(m+1) - 2F(m) \tag{10}$$

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 (9):

$$Angle = \arctan\left(\frac{1}{|F(m)-F(m-1)|}\right) + \arctan\left(\frac{1}{|F(m+1)-F(m)|}\right) \tag{11}$$

In order to select the best clustering validity index, the Angle Based Method (ABM) was performed on the five chosen CVI's. Tables 1 and 2 show the comparison between the method and the process of choosing the first minimum or maximum value depending on the used index. The following procedure is performed to obtain the K estimations:

```
Initialization: Nb_CVI's = 5; k_min = 2; k_max = 20
for i = 1 to number of CVI's (Nb_CVI's) do
for k = k_min to k_max do
    run the K-Harmonic Means Algorithm on labeled data S_i (i = 1...4) with k centres
    compute the value of CVI_i
    end for
select the optimal number of cluster K using the Angle Based Method.
end for
```

EXPERIMENTAL RESULTS AND DISCUSSION

Series of tests were conducted in order to ensure the validity and effectiveness of the proposed method. All the experimental results were obtained using the MATLAB software package.

Comparison between the five 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. Four 2D synthetic data sets were employed during our evaluation. These data sets possess the same number of objects and clusters (N = 5000 objects, K = 15 clusters) with different degrees of overlapping, as depicted in Figure 1. The overlapping allows us to select the optimal CVI that approximates the number of clusters (k = 15) correctly. These data sets are extracted from UCI Repository http://cs.uef.fi/sipu/datasets.

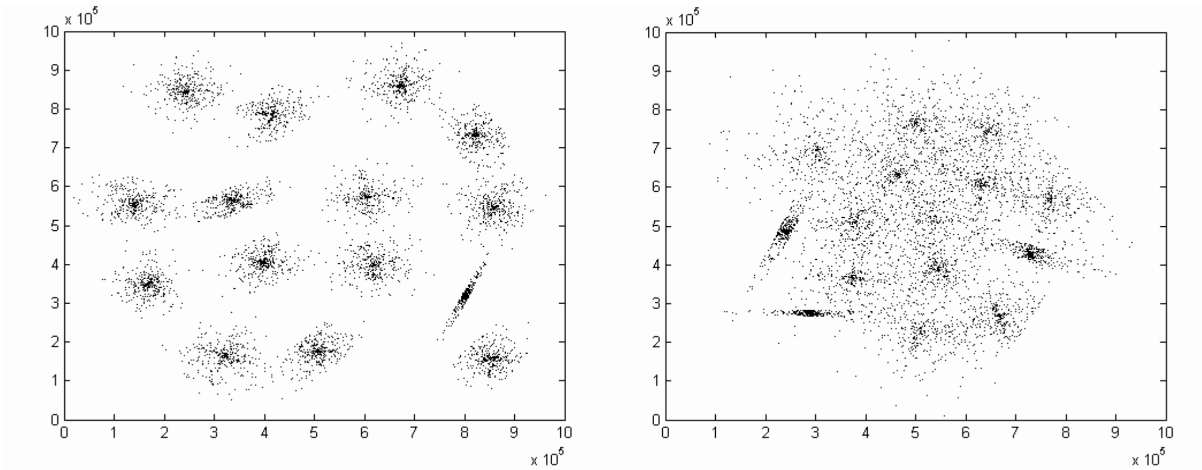

Figure 1: Synthetic data S1 and S4.

Table 1: Comparison among the five CVI's for K-Harmonic Means using S1 data set with 15 clusters.

KHM		**DB**	**XB**	**WB**	**DB***	**BIC**
	***K* without ABM**	5	4	14	2	14
		ADB	**AXB**	**AWB**	**ADB***	**ABIC**
	***K* with ABM**	14	14	14	16	16
	Delta (δ)	0.01	0.2	0.2	10	5
K-Means		**DB**	**XB**	**WB**	**DB***	**BIC**
	***K* without ABM**	7	7	15	2	15
		ADB	**AXB**	**AWB**	**ADB***	**ABIC**
	***K* with ABM**	15	15	15	15	15
	Delta (δ)	0.01	0.2	0.2	10	5

Table 1 illustrates the efficiency of the Angle Based Method in order to find the correct number of clusters. Commonly, the first significant minimum value is selected as the optimal number of clusters as shown in Table 1 for K without ABM. However, the results mentioned above show that the indices are very fluctuant making the returned values inaccurate even knowing that the clusters in the S1 data set well separated. We also notice that both BIC and WB give the correct number of clusters in the case of S1 without using the ABM. Regarding the used algorithms, they delivered approximately the same number of clusters with a small advantage of the K-means algorithm.

Table 2: Comparison among the five CVI's for K-Harmonic Means using S4 data set with 15 clusters.

		DB	**XB**	**WB**	**DB***	**BIC**
KHM	***K* without ABM**	5	5	15	3	3
		ADB	**AXB**	**AWB**	**ADB***	**ABIC**
	***K* with ABM**	16	15	15	19	15
	Delta (δ)	10	0.01	0.01	0.01	5
		DB	**XB**	**WB**	**DB***	**BIC**
K-Means	***K* without ABM**	4	5	15	15	5
		ADB	**AXB**	**AWB**	**ADB***	**ABIC**
	***K* with ABM**	18	15	4	9	15
	Delta (δ)	10	0.01	0.01	0.01	5

Table 2 shows the results for the highly overlapped data set S4. The number of success decreases dramatically when the cluster centres are moved close to each other. 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 to the correct number of clusters, making the use of the ABM 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 due to a bad approximation of the threshold; for example, the WB went from 15 to 4 clusters when applying the ABM. According to the obtained results, we decided to combine the method with the KHM algorithm that gives more accurate estimations in most cases.

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 data sets. Even considering some indices like the WB returns good results, (?)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 data sets. Most of the indices present the same properties in terms of complexity and computing time and give approximately the correct number of clusters. The main reason that made us choose the BIC index is its adaptability among the used data sets and the height improvement by the index while combined with the ABM.

Experiment on Remotely Sensed Data

Besides the synthetic data sets, three sub-scenes acquired by different sensors and given without any ground truth data were applied in the second experiment. The analysis is only based on the visual aspect of the results. The key characteristics of remotely sensed data used in this section are presented in Table 3.

The clustering results of the three images by the proposed method using the three RGB bands are shown in Figure 2d for the Spot-5 sensor, Figure 2e for the Alsat-2A sensor with seven clusters, and Figure 2f for the Landsat 8 sensor 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 urban and cloud pixels, especially in third image. Confusion areas appear because of close radiometric values in the original images that have undergone radiometric corrections. We also notice that shadow effects are reported as a unique cluster, which is also due to the usage of only colorimetric (RGB) values when processing the data.

Table 3: Key characteristics of remotely sensed datasets.

	Size (m^2)	Resolution (m)	Satellite	Area (west of Algeria)	Acquisition date	Preprocessing
Sub-scene 1	400×400	20	Spot-5	Oran	3rd March 2012	Level 2A
Sub-scene 2	500×500	10	Alsat-2A	Tlemcen	4th May 2011	Level 2A
Sub-scene 3	600×800	30	Landsat 8	Arzew	27th Jun 2014	Level L1T

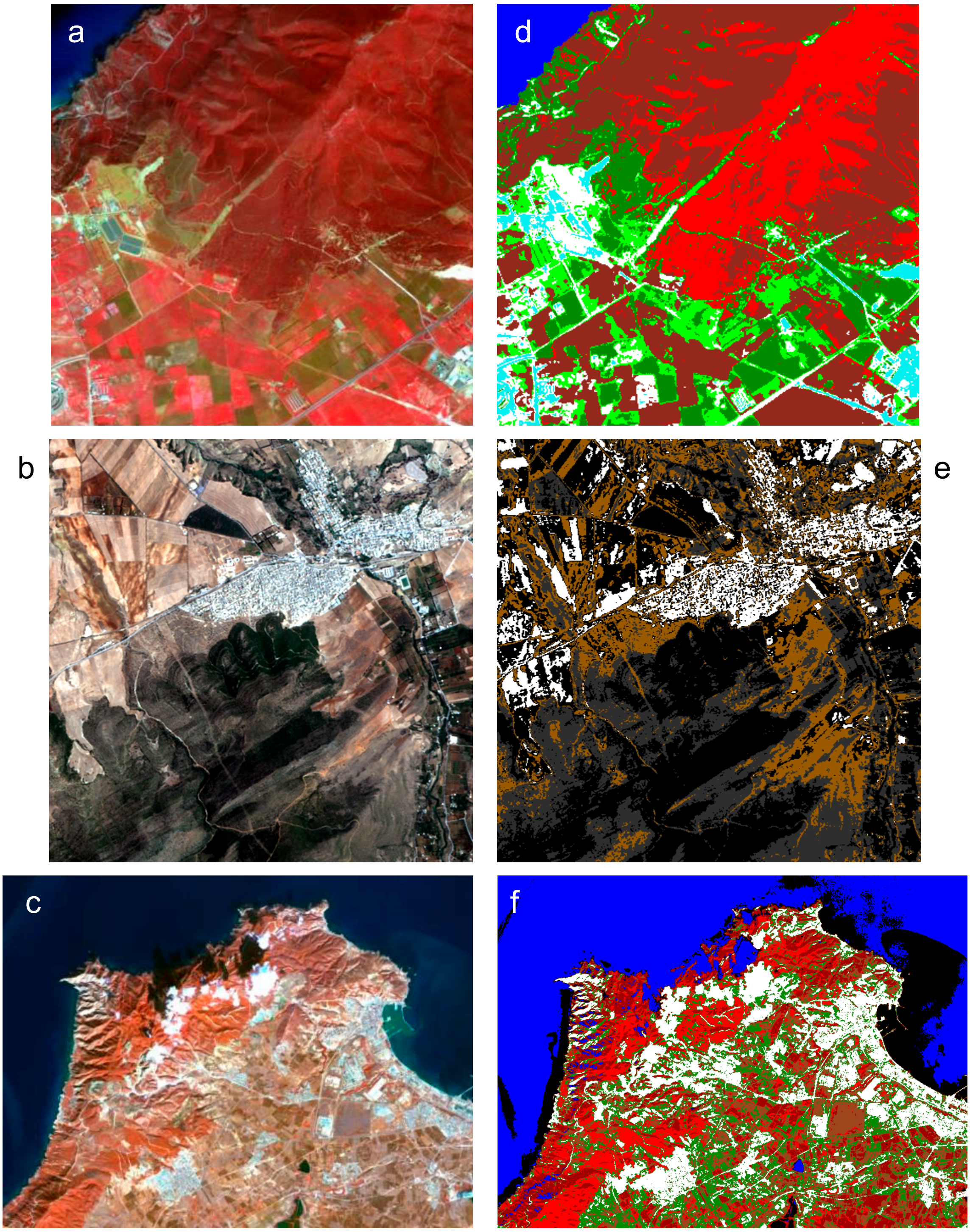

Figure 2: Clustering using the KHM on remotely sensed data sets.

CONCLUSIONS

In this paper, we evaluated the effectiveness of five CVIs on four synthetic data sets and three types of remote sensing data sets by using the KHM and KM algorithms for data set clustering. From the experimental results, it was found that four of the used CVIs failed to return the optimal number of clusters, except the case of the WB index which delivered the right number of clusters. On the other hand, the angle-based method was introduced with the four CVIs to avoid the local optima issues and consequently, to improve the results by returning the accurate number of clusters. Indeed, the results prove the efficiency of the proposed process against using a simple selection method by choosing the first significant minimum value. Additionally, the comparison between the well-known K-means algorithm and the K-Harmonic means shows the superiority of the latter.

Further research will involve the combination of both clusters validity indices and tangle-based method with the Growing KHM (17), which is an improved version of the KHM.

REFERENCES

1 Gan G, C Ma & J Wu, 2007. Data Clustering: Theory, Algorithms, and Applications. ASA-SIAM Series on Statistics and Applied Probability (SIAM, Philadelphia, PA, USA) 466 pp.

2 Jain A K & R C Dubes, 1988. Algorithms for Clustering Data (Prentice-Hall, NJ, USA) 320 pp.

3 MacQueen J, 1967. Some methods for classification and analysis of multivariate observations. In: 5th Berkeley Symposium Mathematics, Statistics and Probability, Vol. 1: Statistics (University of California Press, Berkeley, CA, USA) pp. 281-297

4 Ball G & D Hall, 1965. ISODATA, a Novel Method of Data Analysis and Pattern Classification. Technical report AD 699 616 (Stanford Research Institute, Menlo Park, CA, USA) 79 pp. (last date accessed: 23 Aug 2016)

5 Zhao Q, 2012. Cluster Validity in Clustering Methods. Ph.D. Dissertation, University of Eastern Finland, 189 pp.

6 Gitanjali S K, R R Sedamkar & K Bhandari, 2012. Hyperspectral Image Classification on Decision Level Fusion. International Journal of Computer Applications, IJCA Proceedings on International Conference and Workshop on Emerging Trends in Technology (ICWET 2012) icwet(7): 1-9 (last date accessed: 23 Aug 2016)

7 Zhang B, 2000. Generalized K-Harmonic Means -- Boosting in Unsupervised Learning. Technical Reports, HP Labs Technical Reports, HPL-2000-137, 13 pp.

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

9 Talon J B, S Bourennane, W Philips, D Popescu & P Scheunders (Editors), 2008. Advanced Concepts for Intelligent Vision Systems. Lecture Notes in Computer Science, Vol 5259 (Springer, Berlin, Heidelberg, Germany) 229 pp.

10 Davies D & D Bouldin, 1979. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2): 224-227

11 Thomas J C R, 2013. New version of Davies-Bouldin Index for clustering validation based on cylindrical distance. V Chilean Workshop on Pattern Recognition CWPR 2013, Temuco, Chile, 5 pp. (last date accessed: 23 Aug 2016)

12 Xie X L & A Beni, 1991. Validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 3: 841-846

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

14 Hammerly G & C Elkan, 2002. Alternatives to the K-means algorithm that find better clusterings. Proceedings of the 11th International Conference on Information and Knowledge Management, 600-607

15 Zhang L, L Mao, H Gong & H Yang, 2013. A K-harmonic means clustering algorithm based on enhanced differential evolution. Fifth International Conference on Measuring Technology and Mechatronics Automation, pp. 13-16

16 Thangavel K & K Karthikeyani Visalakshi, 2009. Ensemble based distributed K-harmonic means clustering. International Journal of Recent Trends in Engineering, 2(1): 125-129 (last date accessed: 23 Aug 2016)

17 Mahi H, N Farhi & K Labed, 2015. Remotely sensed data clustering using K-harmonic means algorithm and cluster validity index. Computer Science and Its Applications, 5th IFIP TC 5 International Conference CIIA 2015, Saida, Algeria, Proceedings (Springer), pp. 105-116

18 Akaike H, 1974. A new look at the statistical model selection identification. IEEE Transactions on Automatic Control, AC-19: 719-723