Remotely Sensed Data Clustering Using K-Harmonic Means Algorithm and Cluster Validity Index

Habib Mahi[1(✉)], Nezha Farhi[1], and Kaouter Labed[2]

[1] Earth Observation Division, Centre of Space Techniques, Arzew, Algeria
[2] Kaouter LABED, Faculty of Mathematics and Computer Science Mohamed Boudiaf, University – USTOMB, Oran, Algeria
{hmahi,nfarhi}@cts.asal.dz,
kaouther.labed@univ-usto.dz

Abstract. In this paper, we propose a new clustering method based on the combination of K-harmonic means (KHM) clustering algorithm and cluster validity index for remotely sensed data clustering. The KHM is essentially insensitive to the initialization of the centers. In addition, cluster validity index is introduced to determine the optimal number of clusters in the data studied. Four cluster validity indices were compared in this work namely, DB index, XB index, PBMF index, WB-index and a new index has been deduced namely, WXI. The Experimental results and comparison with both K-means (KM) and fuzzy C-means (FCM) algorithms confirm the effectiveness of the proposed methodology.

Keywords: Clustering · KHM · Cluster validity indices · Remotely sensed data · K-means · FCM

1 Introduction

Clustering is an exploratory data analysis tool that reveals associations, patterns, relationships, and structures in masses of data [1] [2]. Two approaches of clustering algorithms exist in the literature: fuzzy (or soft) and crisp (or hard) clustering. In the first approach, clusters are overlapping and each object belongs to each cluster to a certain degree (or with a certain fuzzy membership level) [3]. The fuzzy c-means (FCM) [4] seems to be the most widely used algorithm in the field of fuzzy clustering. It appears to be an appropriate choice in multiple domains as remote sensing satellite images and pattern recognition [5] [6]. In crisp clustering, clusters are disjoint: each objet belongs to exactly one cluster as example we cite the K-Means (KM) [7] and ISODATA (Iterative Self-Organizing Data Analysis Technique) algorithms [8]. These latter are widely used clustering methods for multispectral image analysis [9]. Also, these algorithms have been successfully used in various topics, including computer vision and astronomy. Their popularity is mainly due to their scalability and simplicity. However, they suffer from a number of limitations. Firstly, the requirement to define a priori the number of K clusters is considered as a handicap and consequently an inappropriate choice of initial clusters may generate poor clustering results [10]. Secondly,

A. Amine et al. (Eds.): CIIA 2015, IFIP AICT 456, pp. 105–116, 2015.
DOI: 10.1007/978-3-319-19578-0_9

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 with clusters which are more or less elongated with a much larger variability, such as forest for example [11]. Also, convergence to local optimum is always observed in this kind of algorithms [1].

To deal with these drawbacks, considerable efforts have been made to mainly create variants from the original methods. As examples we cite KM and its alternatives K-Harmonic Means, Trimmed k-means and k-modes algorithm [1]. At the same time some works have focused on the developing of measures to find the optimal number of clusters using cluster validity indices [3]. We distinguish fuzzy indices (used with fuzzy clustering) and crisp indices (used with hard clustering). As examples of fuzzy indices we can mention XB index [12] as well as Bezdek's PE and PC indices [13] [14]. DB-index [15], Dunn's index [16] and Calinski-Harabasz index [17] are some of the popular indices used in crisp clustering. [3] [18] [19] give a very important review of different CVIs present in the literature.

In this study we investigate the ability of the K-Harmonic Means clustering algorithm combined with validity indices, especially in unsupervised classification of remote sensing data. The rest of paper is organized as follows. Methodology will be firstly presented in Section 2; the experimentation and the results obtained will be tackled in Section 3. Section 4 concludes the paper.

2 Methodology

In this section we give a brief description of the K-Harmonic Means and four clustering validity indices. Then we present the proposed method in details. In the next sections, the following notation will be adopted:

N: The number of objects in the data set.
x_i: The i^{th} object in the data set.
K: The number of clusters.
c_j: The center of cluster j.
d: The number of dataset dimensions.

2.1 K-Harmonic Means Algorithm

The initialization of centers influence on the K-Means (KM) performance and it is considered as the main drawback of this algorithm. To improve KM, Zhang [20] proposes to use the harmonic mean instead of standard mean in the objective function and has named the new algorithm K-Harmonic Means (KHM).

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

New centers clusters are calculated as following [21][22]:

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

2.2 Clustering Validity Indices

In this sub-section, we introduce the clustering validity indices used in this work, namely Davies-Bouldin (DB), Xie-Benie (XB), Pakhira-Bandyopadhyay-Maulik Fuzzy (PBMF), WB index (WB) and WB-XB index (WXI).

- Davies-Bouldin index (DB ↓) [15]: It is a very popular and used crisp index in clustering algorithms. It requires only two parameters to be defined by the user, the distance measure noted p and the dispersion measure noted q. The DB is defined as follows:

$$DB = \frac{1}{K} \sum_{i=1}^{K} R_i \quad (3)$$

With

$$R_i = max_{i,i \neq j} \left\{ \frac{S_i + S_j}{M_{ij}} \right\} \quad (4)$$

Where

$$S_i = \left\{ \frac{1}{T_i} \sum_{j=1}^{T_i} \left\| x_j - c_i \right\|^q \right\}^{\frac{1}{q}} \quad (5)$$

And

$$M_{ij} = \left\{ \sum_{k=1}^{K} \left\| c_{ki} - c_{kj} \right\|^p \right\}^{\frac{1}{p}} \quad (6)$$

With
c_{ki}: k^{th} Component of the n-dimensional vector c_i.
c_i: The center of cluster i.
M_{ij}: The Minkowski metric.
T_i: The number of vectors (pixels) in cluster i.

- Xie-Benie index (XB ↓) [12]: Also called function S, is defined as a ratio of the total variation to the minimum separation of clusters. Its definition is:

$$XB = \frac{1}{N} \frac{\sum_{i=1}^{K} \sum_{j=1}^{N} (\mu_{ij})^m \|x_j - c_i\|^2}{min_{l \neq i} \|c_l - c_i\|^2} \quad (7)$$

- Pakhira-Bandyopadhyay-Maulik Fuzzy index (PBMF ↑) [3]: It is considered as validity index measure for fuzzy clusters. It is formulated as follows:

$$PBMF = \frac{1}{K} \times \frac{E_1}{\sum_{i=1}^{N} \sum_{j=1}^{K} (\mu_{ij})^m \|x_i - c_j\|^2} \times max_{l \neq i} \|c_l - c_i\|^2 \quad (8)$$

With E_1 is constant for a given dataset.

- WB index (WB ↓) [23]: It is defined as a ration of the measure of cluster compactness to its measure of separation. It is given by:

$$WB = K \frac{\sum_{i=1}^{N} \|x_i - c_{pi}\|^2}{\sum_{i=1}^{K} n_i \|c_i - \bar{X}\|^2} \quad (9)$$

- WB-XB index (WXI ↓): It is defined as the average between WB and XB indices and is formulated as follows:

$$\text{WXI} = (\text{WB\_index} + \text{XB\_index}) / 2 \quad (10)$$

2.3 Mean Square Error (MSE)

It is a measure of error which is often used in clustering problems. It represents the mean distance of objects in the dataset from the nearest centers [24]. It is formulated as follows:

$$MSE = \frac{\sum_{j=1}^{K} \sum_{X_i \in C_j} \left\| x_i - c_j \right\|}{N * d} \quad (11)$$

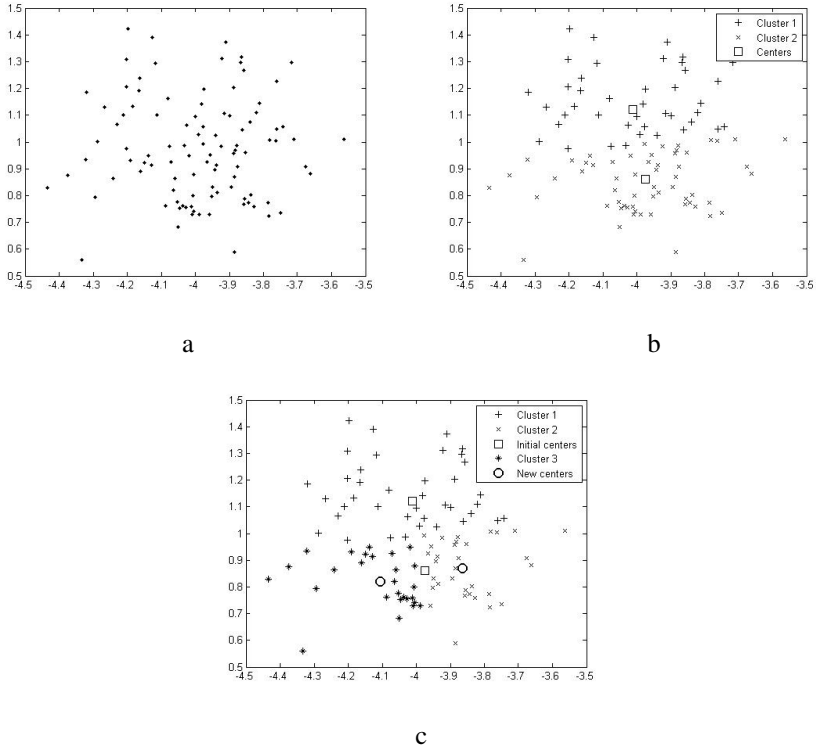
2.4 Proposed Method

In this subsection, we present the proposed method which combines the KHM algorithm, the mean square error (MSE) and WXI cluster validity index. This new method is called Growing KHM (GKHM).

For a given data distribution two centers are chosen randomly (Fig. 1. a), the KHM clustering algorithm is then applied to obtain the two initial clusters (Fig. 1. b). Also, the mean square error (MSE) is computed in this stage for each cluster to select the heterogeneous one (MSE is maximal) to be divided. Therefore, two new centers are computed (Fig. 1. c) and the old one is removed. The process is repeated until a number of epochs are satisfied. The complete algorithm for the proposed method is given by the following:

1. Choose two centers randomly from the dataset.
2. Run the KHM algorithm with these two centers
3. **Repeat**

4. epoch =1
5. Compute MSE for each cluster
6. Select the cluster with the maximum MSE value
    - Insert two new centers halfway between the old center and the two extremes of the cluster in order two have two new clusters.
    - Remove the old center
    - Run the KHM algorithm with the new centers (K = 2).
7. Compute the WXI^{epoch} of all the clusters and save it in the vector V with the related centers
8. **Until** (epoch number's reached)
9. Select the minimum value of WXI in V i.e. The final number of clusters
10. Clustering dataset with the appropriate centers.



a b c

Fig. 1. Process of new centers: a) Data Initialization, b) Data Sampling, c) New Centers Generation

3 Experimental Results

This section is devoted to experiments that ensure the validity and effectiveness of the proposed method. It is divided into three subsections. In the first subsection, an experiment is conducted by using synthetic datasets to select the most suitable cluster validity

index for our work. In the second subsection, a comparison of our approach with both KM and FCM algorithms is drawn. The last subsection concerns the clustering of real satellite images using the proposed method and its results. All the experiments results have been obtained using the MATLAB software package.

3.1 Comparison between the Four Cluster Validity Indices

In order to select the best clustering validity index, we experimentally evaluated their performance on four different synthetic datasets using the basic KHM. Some of the datasets namely S1 and S4 are plotted in Fig. 2 respectively. Each dataset consists of 5000 points representing 15 clusters. All the datasets can be found in the SIPU web page `http://cs.uef.fi/sipu/datasets`.

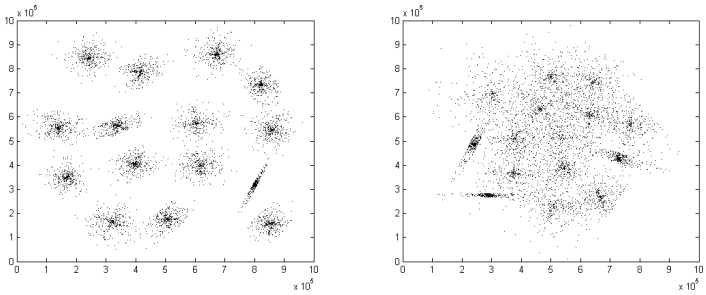


Fig. 2. Synthetic data S1 and S4

In this paper, we have used four synthetic datasets S1-S4 which have the same number of clusters (K=15) and the same Gaussian distribution with increasing overlap between the clusters. The overlapping is an additional criterion which allows us to select the optimal cluster validity index between indices used in this work. For this end, we have applied the KHM algorithm as mentioned before for each dataset by varying the number of clusters from 2 to 20; and, the values of the four CVI's are computed for different K. The results reported in Tables 1 and 2 show only the best values obtained by the four CVI's and their corresponding number of clusters K.

From Table 1, we can see that WB and XB cluster validity indices give the best values for the KHM algorithm and reach their minimum respectively at the optimal

Table 1. Comparison between DB, WB, PBMF and XB indices for S1 dataset

	Cluster Validity Indices			
K	DB	WB	PBMF	XB
13	**0.40**	0,49	2,99 x 10^{10}	0,08
14	0.42	0,36	**9,69 x 10^{10}**	0,06
15	0.44	**0,24**	4,26 x 10^{10}	**0,04**

Table 2. Comparison between DB, WB, PBMF and XB indices for S4 dataset

	Cluster Validity Indices			
K	**DB**	**WB**	**PBMF**	**XB**
4	0.84	1.89	**2.32 x 10^{10}**	0.16
11	**0.64**	1.17	1.18 x 10^{10}	0.11
14	0.65	0.96	1.25 x 10^{10}	**0.09**
15	0.72	**0.90**	0.77 x 10^{10}	0.14

number of clusters (K=15). On the other hand, the DB and PBMF cluster validity indices approximate the number of clusters (K=13 and K=14).

From Table 2, we notice that WB index still offers the best values for the KHM algorithm and reaches its minimum for the optimal clusters number (K=15). The XB index approximate the solution and has its optimum value nearly to the solution (K=14). However, DB and PBMF fail to find a near best solution by returning a completely wrong number of clusters (K=11 and K=4) and having an unstable minimum.

In Summary, the results show that all the cluster validity indices provide an accurate estimation of the clusters number when the clusters in dataset present a small distortion. However, several knee points are detected with exception of the WB index. For clusters with a largest distortion, case of the S3 and S4 datasets, the DB and PBMF indices fail to find the optimal number of clusters. These conclusions lead us to say that only the WB and the XB indices can be used for this kind of datasets. According to the obtained results; the combination of WB and XB indices seems interesting and the new index called WXI (Equation 9) was deduced and tested.

Table 3. Comparison of the minimal values of the WXI for S1,S2,S3 and S4

	S1	**S2**	**S3**	**S4**
OVI	0.13	0.23	0.45	0.49
K	15	15	14	15

The results of WXI are very promising since the error margin reported in Table 3 is acceptable. Indeed, the combination of WB and XB indices has maximized the performances of both of them and erased the deficiency of each one.

3.2 Comparison with KM and FCM Algorithm

In this section, different tests have been performed using GKHM, KM and FCM over 50 iterations. The WXI has been computed in each test and used to compare between their results.

In order to compare between the GKHM and the two other algorithms using the WXI, we have computed up each of them to 50 iterations with a static number of clusters (K=15) for the KM and the FCM. The results appear in Figures 3 and 4.

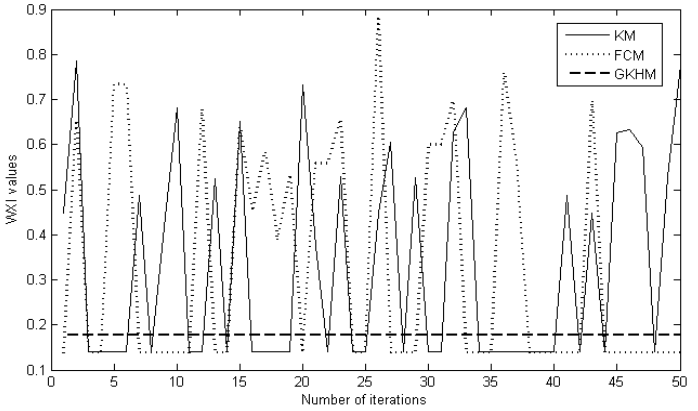


Fig. 3. Comparison between the GKHM, the KM and the FCM for S1 using the WXI

Form Fig. 3, we notice that the KM and the FCM reach inferior minimums than the GKHM but the results are very fluctuant and change constantly; it brutally increases after reaching the minimum which indicates unstable algorithms unlike the GKHM which is totally stable and remains on its minimum value.

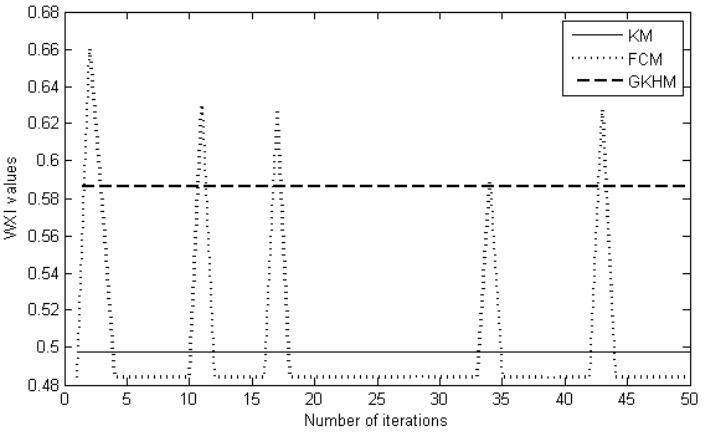


Fig. 4. Comparison between the GKHM, the KM and the FCM for S4 using the WXI

The results in Fig. 5 represent the WXI values for the high overlap data synthetic S4. The curves shape shows that the KM and GKHM are stable; however, only the first algorithm gives the best results. In contrast, the shape of the FCM curve stays very fluctuant and unstable.

Table 4. The average results of the WXI at 50 iterations for all synthetic datasets

	Aver-age_WXI_KM	Average_ WXI _FCM	Average_ WXI _GKHM
S1	0.45	0.43	**0.17**
S2	0.32	0.30	**0.29**
S3	**0.42**	0.45	0.53
S4	**0.49**	0.50	0.64
Average	0.42	0.42	**0.40**

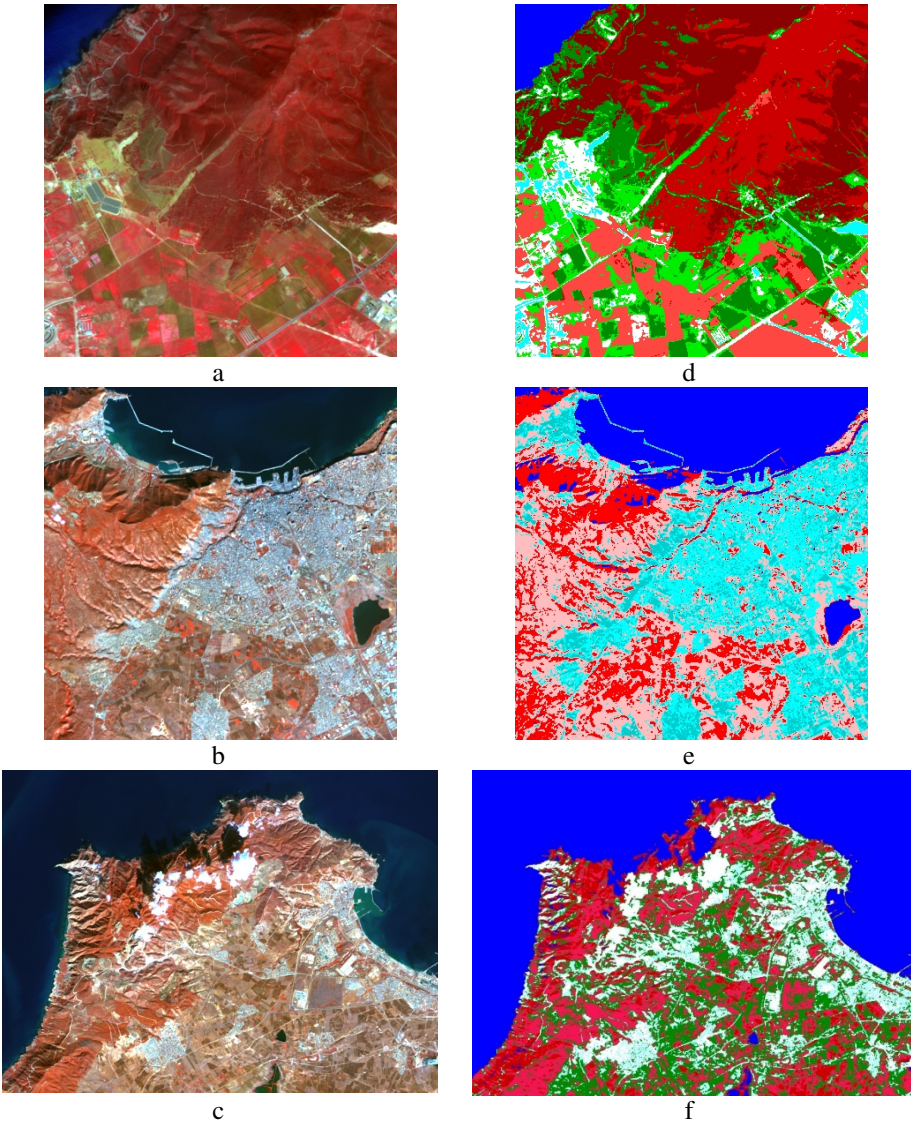


Fig. 5. Clustering using the GKHM on remote sensed data sets

From Table 4, we notice that the GKHM is a totally stable algorithm and tends to minimize the WXI values more than the KM and FCM, especially when data are well-separated. However, the GKHM responds less well when dealing with high overlapped datasets. In the case of FCM and KM, the results are unstable due to their high dependency on their centers number initialization. The three algorithms have approximately the same results with better global issues for GKHM concerning datasets tested in this paper.

3.3 Experiment on Remotely Sensed Data

In the last experiment, the clustering has been performed on three multispectral remotely sensed data; the details of the image sets are as follows:

- A Landsat 8 sub-scene of Oran the image has three spectral data channels and size of 400 x 400. The spatial resolution is 30 meters (Fig. 6.a).
- A Spot 5 sub-scene of Oran the image has three spectral data channels and size of 400 x 400. The spatial resolution is 20 meters (Fig. 6.b).
- A Landsat 8 sub-scene of Arzew the image has three spectral data channels and size of 600 x 800. The spatial resolution is 30 meters (Fig. 6.c).

The clustering results of the three remotely sensed data by the proposed method are shown in Fig. 6.d with eight clusters, Fig. 6.e with five clusters and Fig. 6.f with six clusters, respectively. The visual comparison with the corresponding original images shows that the obtained results appear generally satisfying even if we notice some confusion between water pixels and shadow ones, case of the second image.

4 Conclusion

A new clustering method for multispectral remotely sensed data has been proposed in this paper. The method combines both the K-Harmonic means algorithm and the clustering validity index in order to find the optimal number of clusters and perform the classification task. Note that the K-harmonic means has been used with only two clusters and the increasing of the centers number has been provided by an automatic insertion of the new clusters. However, some improvements can be made, especially by reducing the time processing cycle. Also, the developed algorithm uses internally a combination of validity indices in order to return an optimal number of clusters.

Other improvements could be done by testing the GKHM on large datasets including high-dimensional datasets and shape sets.

A further research will involve the application of new validity indices such as DB* index [25], the comparison with both the enhanced differential evolution KHM [26] and the modified version of k-means algorithm proposed by Celebi and al. [27] and finally the use of the ensemble clustering technique.

References

1. Gan, G., Ma, C., Wu, J.: Data Clustering: Theory, Algorithms, and Applications. ASA-SIAM Serieson Statistics and Applied Probability. SIAM, Philadelphia (2007)
2. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice-Hall, Englewood (1988)
3. Pakhira, M.K., Bandyopadhyay, S., Maulik, U.: A Study of Some Fuzzy Cluster Validity Indices, Genetic Clustering and Application to Pixel Classification. Fuzzy Sets and Systems 155, 191–214 (2005)
4. Bezdeck, J.C.: FCM: Fuzzy C-Means algorithm. Computers and Geoscience 10, 191–203 (1984)
5. Gong, X.-J., Ci, L.-L., Yao, K.-Z.: A FCM algorithm for remote-sensing image classification considering spatial relationship and its parallel implementation. In: International Conference on Wavelet Analysis and Pattern Recognition, ICWAPR 2007, November 2-4, vol. 3, pp. 994–998 (2007)
6. Gao, Y., Wang, S., Liu, S.: Automatic Clustering Based on GA-FCM for Pattern Recognition. In: Second International Symposium on Computational Intelligence and Design, ISCID 2009, December 12-14, vol. 2, pp. 146–149 (2009)
7. McQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proc. 5th Berkeley Symp. Mathematics, Statistics and Probability, pp. 281–296 (1967)
8. Ball, G., Hall, D.: ISODATA: A novel method of data analysis and pattern classification. In Technical report, Stanford Research Institute, Menlo Park, CA, USA (1965)
9. Huang, K.: A Synergistic Automatic Clustering Technique (Syneract) for Multispectral Image Analysis. Photogrammetric Engineering and Remote Sensing 1(1), 33–40 (2002)
10. Zhao, Q.: Cluster validity in clustering methods. Ph.D. dissertation. University of Eastern Finland (2012)
11. Korgaonkar, G.S., Sedamkar, R.R., KiranBhandari.: Hyperspectral Image Classification on Decision level fusion. In: IJCA Proceedings on International Conference and Workshop on Emerging Trends in Technology, vol. 7, pp. 1–9 (2012)
12. Xie, X.L., Beni, A.: Validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell. 3, 841–846 (1991)
13. Bezdek, J.C.: Cluster validity with fuzzy sets. J. Cybernet. 3, 58–73 (1974)
14. Bezdek, J.C.: Mathematical models for systematics and taxonomy. In: Eighth International Conference on Numerical Taxonomy, San Francisco, CA, pp. 143–165 (1975)
15. Davies, D., Bouldin, D.: A cluster separation measure. IEEE PAMI 1(2), 224–227 (1979)
16. Dunn, J.C.: A fuzzy relative of the isodata process and its use in detecting compact well separated clusters. J. Cybernet. 3, 32–57 (1973)
17. Calinski, R.B., Harabasz, J.: Adendrite method for cluster analysis. Commun. Statist. 1–27 (1974)
18. Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Prez, J.M., Perona, I.: An extensive comparative study of cluster validity indices. Pattern Recognition 46(1), 243–256 (2013)
19. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Clustering validity checking methods: Part II. SIGMOD Record 31(3), 19–27 (2002)
20. Zhang, B.: Generalized K-Harmonic Means Boosting in Unsupervised Learning. Technical Reports, Hewllet Laborotories, HPL-2000-137 (2000)
21. Zhang, L., Mao, L., Gong, H., Yang, H.: A K-harmonic Means Clustering Algorithm Based on Enhanced Differential Evolution. In: 2013 Fifth International Conference on Measuring Technology and Mechatronics Automation, 2014 Sixth International Conference on Measuring Technology and Mechatronics Automation, pp. 13–16 (2013)

22. Thangavel, K., Karthikeyani Visalakshi, K.: Ensemble based Distributed K- Harmonic Means Clustering. International Journal of Recent Trends in Engineering 2(1), 125–129 (2009)
23. Zhao, Q., Fränti, P.: WB-index: a sum-of-squares based index for cluster validity. Knowledge and Data Engineering 92, 77–89 (2014)
24. Malinen, M.I., Mariescu-Istodor, R., Fränti K-means*, P.: Clustering by gradual data transformation. Pattern Recognition 47(10), 3376–3386 (2014)
25. Thomas, J.C.R.: New Version of Davies-Bouldin Index for Clustering Validation Based on Cylindrical Distance. In: V Chilean Workshop on Pattern Recognition, November 11-15 (2013)
26. Zhang, L., Mao, L., Gong, H., Yang, H.: A K-harmonic Means Clustering Algorithm Based on Enhanced Differential Evolution. In: 2013 Fifth International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), January 16-17, pp. 13–16 (2013), doi:10.1109/ICMTMA.2013.1
27. Emre, C.M., Kingravi, H.A., Vela, P.A.: A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications (2013)