QALAAI ZANIST JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil – Kurdistan – Iraq
Vol. (2), No. (2), April 2017
Speical Issue : The 1st International Conference on Information Technology (ICoIT'17)
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

A Survey On Unsupervised Evaluation Criteria For Image Clustering Validation

Akar Taher
Department of Software Engineering, Faculty of Engineering, Koya University – University Park, Kurdistan Region of Iraq *(Part time Lecturer at Lebanese French University)*
Akar.taher@koyauniversity.org

ARTICLE INFO

Article History:

Received: 13 March 2017

Accepted: 1 April 2017

Published: 10 April 2017

DOI:

10.25212/lfu.qzj.2.2.10

Keywords: *Evaluation, Criteria, Validation Unsupervised, Supervised, Clustering.*

ABSTRACT

The evaluation of clustering results is the most difficult and frustrating part of cluster analysis. The challenge is to validate the obtained results without any apriori information. Validity indexes are widely used approach for evaluation of clustering results. These approaches can use three criteria: i) external (also called supervised) criteria: this type is based on comparing the obtained results with a previously known result (frequently called ground truth) and compute the similarity, ii) internal criteria (also called unsupervised) criteria: estimate the quality of the result using internal information of the data alone, and iii) relative criteria: this means multiple usages of one of the two above types ofdifferent results and see which is better than the other. Therefore we can say: depending on the information available and the problem type, different types of indexes might be used for cluster validation. Sometimes due to the complexity of the datasets, one validity index is not sufficient to evaluate the quality of the obtained results, and then a combination of two or more index should be used. In this paper, a basic general review on evaluation criteria is first given and then the focus is spotted on unsupervised criteria as they are much more useful, thanks to their objective functionality.

1. INTRODUCTION

The evaluation of clustering [1]–[3] results is an unavoidable process [4]–[7] which is used to quantify the performance of clustering algorithms. The quality evaluation of a clustering result is an active area of research and many criteria are being developed regularly. Unfortunately, the evaluation of a clustering result always contains some elements of subjectivity and the criteria do not always give satisfactory evaluation [8]. For this reason, it is impossible to define a universal criterion to evaluate the results produced by all the existing criteria. However, a number of criteria exist and are repeatedly used by many researchers to compare clustering results [9], [10]. Since there are a large number of possible clustering results for the same dataset, the objective is to assess whether any of these results is better than another. So to correctly evaluate and validate the results, it can be necessary to use multiple evaluation criteria like in [11].

In this paper, the generalities of the evaluation criteria are provided, and then focus is on the unsupervised ones.

This following of this paper is organized as: the second section will describe evaluation criteria types, then followed by unsupervised evaluation criteria in the third section, and fourth section will give the conclusions.

2. EVALUATION CRITERIA TYPES

Several types of evaluation methods have been proposed in the literature [7], [12]. They are classified into three main groups. The first group contains unsupervised criteria that use only internal information of the data such as the distance between objects. These criteria are also called internal quality measures. The second group contains supervised criteria that calculate the degree of correspondence between the clustering produced by the algorithm and a known data partitioning. These criteria are also known as external quality measures. The last group is called relative criteria; this type of evaluation allows comparing the results obtained from the same algorithm. These measures are simply the use of internal or external criteria to evaluate multiple results produced by the same algorithm and to choose the best one among them. In this paper only internal quality criteria are reviewed.

3. UNSUPERVISED EVALUATION CRITERIA

Unsupervised evaluation criteria [9] are based on internal information and do not need any *a priori* knowledge. This type of criteria generally computes statistical measures such as the standard deviation or the disparity of the classes. These measures are often based on the simplest definition of partitioning which says that objects from the same class should be as close as possible, and that objects from two distinct classes should be as far apart as possible [13] (see **Figure 1**) . To assess whether a clustering result complies with this intuitive definition, the distances between the class centers and the class objects are calculated. These unsupervised measures assess the compactness and the separateness of the classes. The evaluation of the quality of a cluster is not formally defined, so there are many different criteria, which estimate the quality of the results differently. Some of these criteria can be directly used as the objective function of a clustering algorithm. However others are very time-consuming, and therefore intended to be calculated after the application of the algorithm for the final evaluation process.

One of the most basic and intuitive criteria able to quantify the quality of a clustering result is the within-class uniformity. The simplest way to calculate this uniformity is the sum of the squared errors (SSE) which is calculated as follows in Equation (1):

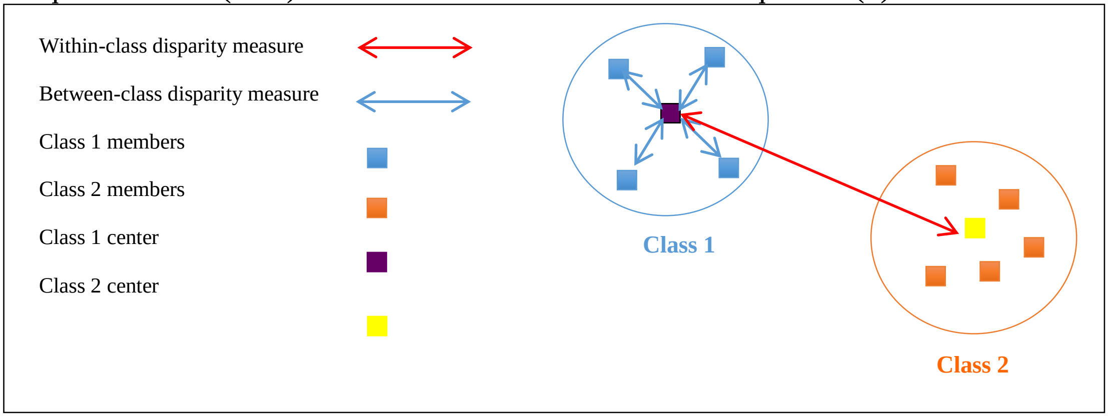


Figure 1: illustration of within-class homogeneity and the between-class disparity

$$SSE\left(I_R\right)=\sum_{i=1}^{NC}\sum_{x\in c_i}d\left(x-g(C_i)\right)^2 \qquad (1)$$

where $g(C_i)$ is the center of the class C_i and d is a distance measure.

Weszka and Rosenfeld [14] proposed such a criterion with thresholding that measures the effect of noise to evaluate some thresholded images. Based on the same idea of within-class uniformity, Levine and Nazif [15] also defined a criterion that calculates the uniformity of a class as follows:

$$LEV1\left(I_R\right)=1-\frac{1}{N}\sum_{i=1}^{NC}\frac{\sum_{x\in C_i}\left[g_l\left(x\right)-\sum_{x\in C_i}g_l\left(x\right)\right]^2}{\left(max_{x\in C_i}\left(g_l\left(x\right)\right)-min_{x\in C_i}\left(g_l\left(x\right)\right)\right)^2} \qquad (2)$$

where

- I_R is the partitioning result of the image *I* into NC classes $C=\left\{C_1,\ldots,C_{NC}\right\}$,
- N is the number of pixels of the image *I*,
- $g_l\left(x\right)$ is the gray level of pixel *x* in the image *I*.

A standardized uniformity measure was proposed by Sezgin and Sankur [16] that is based on the Cochran homogeneity measurement [17]. However, this method requires a threshold that is often arbitrarily selected, thus limiting the usage of this criterion. Another criterion to measure the within-class uniformity was developed by Pal and Pal [18]. It is based on a thresholding that maximizes the local entropy of the classes in a partitioning result. In the case of slightly textured images, these criteria of within-class uniformity prove to be effective and very simple to use. However, the presence of textures in an image often generates improper results due to the over-influence of small regions.

Complementary to the within-class uniformity, Levine and Nazif [15] defined a disparity measure between two classes to evaluate the dissimilarity of different classes in a partitioning result. The formula of total between-class disparity is defined as follows:

$$LEV2\left(I_R\right)=\frac{\sum_{k=1}^{NC}w_{C_k}\sum_{j=1/C_j\in W\left(C_k\right)}^{NC}\left[p_{C_k\backslash C_j}\left(\left|g_l\left(C_k\right)-g_l\left(C_j\right)\right|/\left(\left(g_l\left(C_k\right)+g_l\left(C_j\right)\right)\right)\right)\right]}{\sum_{k=1}^{NC}w_{R_k}} \qquad (3)$$

where w_{c_k} is a weight associated to C_k that can be dependent of its area, $g_l\left(C_k\right)$ is the average of the gray level of C_k and $p_{C_k\backslash C_j}$ is the length of the boundary of the class C_k common to the perimeter of the class C_j. This type of criterion has the advantage of penalizing over-segmentation.

Zeboudj [19] proposed a measure based on the combined principles of maximum between-class (external) disparity and minimal within (interior) class disparity measured at the pixel's neighborhood.

Let $c(x,z)=\dfrac{|g_l(x)-g_l(z)|}{L-1}$ be the disparity between two pixels x and z x and $z \in C_i$, and L be the maximum gray level.

The interior disparity $CI(C_i)$ of the class C_i is defined as follows:

$$CI(C_i)=\frac{1}{NC_i}\sum_{x\in C_i}\max\left\{c(x,z),\ z\in\ V_s,\ z\in C_i\right\} \qquad (4)$$

where NC_i is the number of pixels in class C_i and V_s is the neighborhood of the pixel x.

The external disparity $CE(C_i)$ of the class C_i is defined as follows:

$$CE(C_i)=\frac{1}{p_i}\sum_{x\in C_i}\max\left\{c(x,z), z\in V_s,\ z\notin C_i\right\} \qquad (5)$$

where p_i is the length of the boundary of class C_i .

Lastly, the disparity of the class C_i is defined by the measurement $D(C_i)$ ∈ [0, 1] expressed as follows:

$$D(C_i)=\begin{cases}1-\dfrac{CI(C_i)}{CE(C_i)} & if\ \ 0<CI(C_i)<CE(C_i)\\ CE(C_i) & if\ \ CI(C_i)=0\\ 0 & otherwise\end{cases} \qquad (6)$$

Zeboudj's criterion is defined by:

$$ZEB(I_R)=\frac{1}{N}\sum_{i=1}^{NC}NC_i x D(C_i) \qquad (7)$$

where N is the number of pixels in the image.

This criterion has the disadvantage of not correctly taking into account strongly textured regions.

Another criterion that is based on the combination of the within-class and between-class disparities is the Davies-Bouldin index [20]. It estimates the within-class disparity based on the distance from the points in a class to its centroid and the between-class disparity based on the distance between centroids. It is defined as:

$$DB(I_R)=\frac{1}{NC}\sum_{i=1}^{NC}\max_{j,j\neq i}\left\{\frac{\frac{1}{NC_i}\sum_{k=1}^{NC_i}d\left(F_k,g(C_i)\right)+\frac{1}{NC_j}\sum_{k=1}^{NC_j}d\left(F_k,g(C_j)\right)}{d\left(g(C_i),g(C_j)\right)}\right\} \qquad (8)$$

where F_k is vector of Nf features representing the pixel x_k.

Another criterion of this type is the Silhouette index [21]. This index is a normalized summation-type index. The within-class is measured based on the distance between all the points in the same cluster and the separation is based on the nearest neighbor distance.

Let $d_1\left(x_j\right)$ be the average dissimilarity of x_j with all other pixels of its class C_i $d_1\left(x_j\right)$ indicates how well x_j is assigned to its class (the smaller the value, the better the assignment).

Let $d_2\left(x_j\right)$ be the lowest average dissimilarity of x_j to any other class C_l $with\, l=1,2,...,K;\ l\neq i$.

The class with the lowest average dissimilarity is said to be the "neighboring cluster" of x_j because it is the next best-fit class for it; and then the size of the silhouette $Sil(x_j)$ is defined as:

$$Sil(x_j)=\frac{d_2\left(x_j\right)-d_1\left(x_j\right)}{max\left[d_2\left(x_j\right),d_1\left(x_j\right)\right]} \tag{9}$$

Basing on the definition of $Sil(x_j)$, the silhouette of the class C_i is defined as:

$$Sil(C_i)=\frac{1}{NC_i}\sum_{x_j\in C_i}Sil(x_j) \tag{10}$$

Finally the global silhouette for a partition is defined as:

$$Sil\left(I_R\right)=\frac{1}{NC}\sum_{i=1}^{NC}sil(C_i) \tag{11}$$

This criterion is very efficient but its time complexity makes it inapplicable to large datasets.

The Dunn’s index (*Du*) [22] is another unsupervised criterion that measures the compactness of a class and the separateness between classes as follows:

$$Du\left(I_R\right)=\frac{\min\limits_{i=1:NC}\left(\min\limits_{j=1:NC,j\neq i}\left(d_{ij}\left(g(C_i),g(C_j)\right)\right)\right)}{\max\limits_{i=1:NC}\left(d_{ii}(g(C_i),g(C_i))\right)} \tag{12}$$

where $d_{ij}\left(g(C_i),g(C_j)\right)$ is the distance between the center of classes $C_i\, and\, C_j$, which is defined here as the minimum distance between the objects of different classes (see Equation (13)). $d_{ii}(g(C_i),g(C_i))$ is the maximum distance between two objects in the same class (see Equation (14)).

$$d_{ij}\left(g(C_i),g(C_j)\right)=\min_{x\in C_i,y\in C_j}d(x,y) \tag{13}$$

$$d_{ii}\left(g(C_i), g(C_i)\right) = \max_{x,y\in C_i} d(x,y) \tag{14}$$

This evaluation criterion has two disadvantages: firstly, it is very time consuming and secondly it is highly affected by the presence of noise in the dataset.

In [23], [24] Rosenberger and Chehdi presented a criterion that enables estimating the within-class homogeneity and the between-class disparity considering the types of regions (textured or non-textured)[1] in the partitioning result. This criterion quantifies the quality of a partitioning result as follows:

$$ROS\left(I_R\right) = \frac{1+\bar{D}\left(I_R\right)-\underline{D}\left(I_R\right)}{2} \tag{15}$$

The global within-class disparity $\underline{D}\left(I_R\right)$ quantifies the homogeneity of each class obtained in the partitioning result I_R of image I. On the other hand, the global between-class disparity $\bar{D}\left(I_R\right)$ quantifies how well the classes obtained are separated from each other.

The global within-class disparity $\underline{D}\left(I_R\right)$ reflects the statistical stability of each class. It is calculated from the within-class disparity $\underline{D}\left(C_i\right)$ of the different classes in a partitioned image:

$$\underline{D}\left(I_R\right) = \frac{1}{NC}\sum_{i=1}^{NC}\frac{NC_i}{N}\underline{D}\left(C_i\right) \tag{16}$$

The weight of the within-class disparity of a class C_i in the global within-class disparity is proportional to the number of pixels for this class. The same principle is used to calculate the between-class disparity $\bar{D}\left(I_R\right)$ of the partitioned image I_R that measures the disparity of each class with the other classes:

$$\bar{D}\left(I_R\right) = \frac{1}{NC}\sum_{i=1}^{NC}\frac{NC_i}{N}\bar{D}\left(C_i\right) \tag{17}$$

This criterion is calculated using the between-class disparity $\bar{D}\left(C_i\right)$ and the with-in class disparity $\underline{D}\left(C_i\right)$ of each class C_i. The calculation of these two criteria is detailed in the following:

– *Within-class disparity criterion*

This criterion evaluates the homogeneity of a class, i.e. the variation of the statistics in the interior of this class. In the calculation of the within-class disparity, the nature of the regions (i.e. textured and non-textured) is taken into account.

In the non-textured case, this criterion for class C_i is defined as:

$$\underline{D}\left(C_i\right) = \sqrt{\frac{1}{NC_i}\sum_{x\in C_i} g_l\left(x\right)^2 - \frac{1}{NC_i^2}\left(\sum_{x\in C_i} g_l\left(x\right)\right)^2} \tag{18}$$

[1] For further reading about region type in an image please refer to [25].

This criterion is sufficient to characterize the within-class disparity of a non-textured region. However, in the textured case, each class is characterized by a set of texture feature vectors. The dispersion of this set of vectors allows calculating the within-class disparity in the textured case.

– *Between-class disparity criterion*

The evaluation process of between-class disparity of a class is similar to the with-in class disparity, but instead of estimating the homogeneity of a class, it is disparity with the other classes is calculated. The between-class disparity is also calculated according to the nature of the regions as follows:

- Between classes of the same region type:
  - The disparity between two classes belonging to uniform regions $\bar{D}(C_i, C_j)$ is defined as:

$$\bar{D}(C_i, C_j) = \frac{|g_I(C_i) - g_I(C_j)|}{NG} \qquad (19)$$

where *NG* is the number of the gray levels in the image

  - The disparity between two classes belonging to textured regions $\bar{D}(C_i, C_j)$ is defined as:

$$\bar{D}(C_i, C_j) = \frac{d\left(g(c_i), g(c_j)\right)}{\|g(c_i)\| + \|g(c_j)\|} \qquad (20)$$

where *d(.,.)* is the Euclidean distance, *g*(*C*$_i$) is the centroid of class C_i, and $\| \, . \, \|$ denotes the Euclidean norm.

- Between classes of different region types: the disparity between classes of different region types is set as the maximum value, i.e. 1.

Error! Reference source not found. gives a summery of the previously cited methods in this section.

Table 1: Summary of main internal (unsupervised) evaluation criteria

Evaluation Criteria	Remarks
Sum of squared errors	Measures within-class disparity.
Levine and Nazif (*LEV1*) [15]	Measures within-class disparity.
Levine and Nazif (*LEV2*) [15]	Measures within class and between class disparities.
Zeboudj index [19]	Measures within class and between class disparities.
Davies-Bouldin index [20]	Measures within class and between classes, time consuming.
Silhouette index [21]	Measures with-in class and between classes, very time consuming.
Dunn index [22]	Measures with-in class and between class disparities, not effective in case of noisy images.

Rosenberger and Chehdi [23], [24]	Measures with-in class and between class disparities, takes into account the region type (textured and non textured)

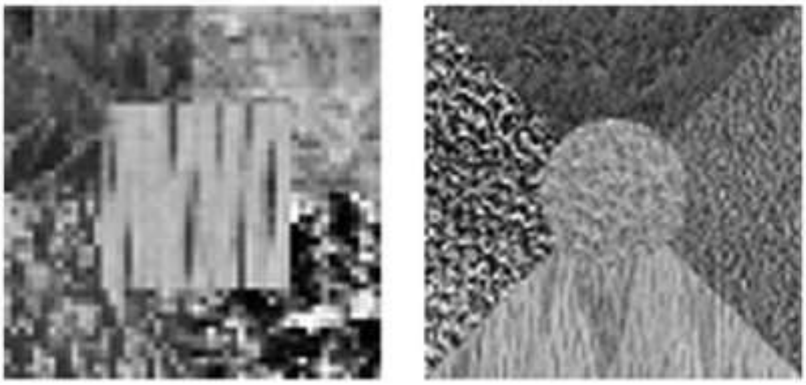
4. EXPERIMENTAL RESULTS

The evaluation criteria described in the previous section are tested on an image database containing 100 images. These images are clustered by the well-known *k*-means, and fuzzy C-means [26] algorithms. As these algorithms are unstable algorithm, the algorithms are run 100 times on the same image and then the result that has the best objective function value of that algorithm is selected. In this database the images are composed of five regions (extracted from the Brodatz album [27]). The number of classes *k* is set to 5, as the images contain 5 regions for both algorithms, and the fuzzification factor *m* is set to 4 to get more stable results [28].

Figure 2 shows two examples of one synthetic image of this database. The unsupervised evaluations are compared to Overall Correct Classification Rate (OCCR), which compares the clustering results with the ground truth (GT) of the image that is calculated as below:

$$OCCR = \frac{No.\ of\ \ pixels\ \ classified\ \ correctly}{Total\ \ No.\ of\ \ Pixels} \qquad (21)$$

According to the results, the Rosenberger and Chehdi index is the best index that in the 88% of the cases the index scores high values and high OCCR. The Silhouette index is the second best index with a score of 79%. Davies-Bouldin and Dunn gave 74% and 72% respectively. The remaining Zeboudj and the 2 indices of Levine and Nazif gave scores below 45%.



(a) (b)

Figure 2: Examples of images used for performance comparison:
(a) Original image 1 and (b) Original image 2.

5. CONCLUSION

The evaluation of clustering results is inevitable to assess the quality of the results obtained.

In this paper, various unsupervised evaluation criteria used to assess the quality of a clustering result is presented. These criteria are also called internal criteria because they do not use any external information in the evaluation process. Each of these criteria are effective in specific cases, for example in image datasets, some are effective the case of non-textured or slightly textured images, while others give effective results in the case of textured images. None of the evaluation methods can prove satisfactory in all the cases. Therefore, to correctly evaluate the algorithms and their results; more than one evaluation technique should be used and their results should be combined.

6. ACKNOWLEDGEMENT

Many thanks goes to the team of TSI2M in the engineering school of ENSSAT in names (Prof. Dr. Kacem Chedi and Asst. Dr. Claude Cariou), Rennes 1 University, France, as this paper is a part of my work in corporation with them.

7. REFERENCES

[1] R. Xu and I. Wunsch, D., “Survey of clustering algorithms,” IEEE Trans. Neural Netw., vol. 16, no. 3, pp. 645–678, 2005.

[2] C. C. Aggarwal and C. K. Reddy, Data Clustering: Algorithms and Applications. CRC Press, 2013.

[3] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data Clustering: A Review,” ACM Comput Surv, vol. 31, no. 3, pp. 264–323, Sep. 1999.

[4] A. Biswas and B. Biswas, “Defining quality metrics for graph clustering evaluation,” Expert Syst. Appl., vol. 71, pp. 1–17, Apr. 2017.

[5] F. Zaidi, D. Archambault, and G. Melançon, “Evaluating the Quality of Clustering Algorithms Using Cluster Path Lengths,” in Advances in Data Mining. Applications and Theoretical Aspects, 2010, pp. 42–56.

[6] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “Cluster Validity Methods: Part I,” SIGMOD Rec, vol. 31, no. 2, pp. 40–45, Jun. 2002.

[7] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “Clustering Validity Checking Methods: Part II,” SIGMOD Rec, vol. 31, no. 3, pp. 19–27, Sep. 2002.

[8] W. M. Rand, “Objective Criteria for the Evaluation of Clustering Methods,” J. Am. Stat. Assoc., vol. 66, no. 336, pp. 846–850, 2012.

[9] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, “Validity index for crisp and fuzzy clusters,” Pattern Recognit., vol. 37, no. 3, pp. 487–501, Mar. 2004.

[10] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, “An extensive comparative study of cluster validity indices,” Pattern Recognit., vol. 46, no. 1, pp. 243–256, Jan. 2013.

[11] K. Chehdi, A. Taher, and C. Cariou, “Stable and unsupervised fuzzy C-means method and its validation in the context of multicomponent images,” J. Electron. Imaging, vol. 24, no. 6, p. 061117, Dec. 2015.

[12] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

[13] R.M. Haralick and L. G. Shaprio, “Image Segmentation Techniques,” presented at the Computer Vision Graphics and Image Processing, Arlington, 1985, vol. 29, pp. 100–132.

[14] J. S. Weszka and A. Rosenfeld, “Threshold Evaluation Techniques,” IEEE Trans. Syst. Man Cybern., vol. 8, no. 8, pp. 622–629, Aug. 1978.

[15] M. D. Levine and A. M. Nazif, “Dynamic Measurement of Computer Generated Image Segmentations,” IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-7, no. 2, pp. 155–164, Mar. 1985.

[16] B. S. Mehmet Sezgin, “Survey over image thresholding techniques and quantitative performance evaluation,” J. Electron. Imaging, vol. 13, pp. 146–168, 2004.

[17] W. G. Cochran, “Some Methods for Strengthening the Common X 2 Tests,” Biometrics, vol. 10, no. 4, p. 417, Dec. 1954.

[18] N. R. Pal and S. K. Pal, “Entropic thresholding,” Signal Process., vol. 16, no. 2, pp. 97–108, Feb. 1989.

[19] R. Zéboudj, Filtrage, seuillage automatique, contraste et contours: du pré-traitement à l’analyse d’image. Saint-Etienne, 1988.

[20] D. L. Davies and D. W. Bouldin, “A Cluster Separation Measure,” IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[21] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” J. Comput. Appl. Math., vol. 20, pp. 53–65, Nov. 1987.

[22] J. C. Dunn†, “Well-Separated Clusters and Optimal Fuzzy Partitions,” J. Cybern., vol. 4, no. 1, pp. 95–104, Jan. 1974.

[23] C. Rosenberger, “Mise en oeuvre d’un systeme adaptatif de segmentation d’images,” Phd Thesis, University of Rennes 1, 1999.

[24] C. Rosenberger, K. Chehdi, and C. Kermad, “Adaptive segmentation system,” presented at the 5th International Conference on Signal Processing, WCCC-ICSP, 2000, vol. 2, pp. 918–921 vol.2.

[25] A. Taher, K. Chehdi, and C. Cariou, “Hyperspectral image segmentation using a cooperative nonparametric approach,” presented at the Image and signal processing for remote sensing XIX, Dresden-Germany, 2013, vol. 8892, p. 88920J–88920J–8.

[26] V. K. Dehariya, S. K. Shrivastava, and R. C. Jain, “Clustering of Image Data Set Using K-Means and Fuzzy K-Means Algorithms,” presented at the International Conference on Computational Intelligence and Communication Networks (CICN), 2010, pp. 386–391.

[27] P. Brodatz, Textures: A Photographic Album for Artists and Designers. Dover Publications, Incorporated, 1999.

[28] A. Taher, Approche coopérative et non supervisée de partitionnement d’images hyperspectrales pour l’aide à la décision. Rennes 1, 2014.