Optimal Fuzzy Clustering in Overlapping Clusters

Ouafa Ammor[1], Abdelmonaime Lachkar[2], Khadija Slaoui[3], and Noureddine Rais[1]
[1] Department of Mathematics, Faculty of Sciences and Technology of Fes, Morocco
[2] ESTM, Moulay Ismail University, Morocco
[3] Department of Physics, Faculty of Sciences Dhar Mehraz of Fes, Morocco

Abstract: *The fuzzy c-means clustering algorithm has been widely used to obtain the fuzzy k-partitions. This algorithm requires that the user gives the number of clusters k. To find automatically the "right" number of clusters, k, for a given data set, many validity indexes algorithms have been proposed in the literature. Most of these indexes do not work well for clusters with different overlapping degree. They usually have a tendency to fails in selecting the correct optimal clusters number when dealing with some data sets containing overlapping clusters. To overcome this limitation, we propose in this paper, a new and efficient clusters validity measure for determination of the optimal number of clusters which can deal successfully with or without situation of overlapping. This measure is based on maximum entropy principle. Our approach does not require any parameter adjustment, it is then completely automatic. Many simulated and real examples are presented, showing the superiority of our measure to the existing ones.*

Keywords: *Unsupervised clustering, cluster validity index, optimal clusters number, overlapping clusters, maximum entropy principle.*

Received November 30, 2006; accepted June 12, 2007

1. Introduction

Cluster analysis has been playing an important role in solving many problems in pattern recognition, image processing, colour image segmentation, machine learning, data mining, and different fields like medicine, biology, technology, marketing. The aim of the cluster validity is to find the partitioning that best fits the underlying data. A wide variety of clustering algorithms have been proposed for different applications and a good overview can be found in the literature [9, 14, 16, 17].

Since there are no predefined classes, it is therefore difficult to find an appropriate metric for measuring if the found clusters configuration is acceptable or not. The result of a clustering algorithm can be very different from each other on the same data set, and input parameters of an algorithm can extremely modify the behaviour and execution of that algorithm. Usually, in well separated clusters, 2D data sets are used for evaluating clustering algorithms as the reader can easily verify the result. But in case of high dimensional data, the visualization and visual validation is not a trivial task. Therefore some formal methods are needed.

The process of evaluating the results of a clustering algorithm is called cluster validity assessment. For this, there are three different techniques: external criteria, internal criteria and relative criteria. Both internal and external criteria are based on statistical methods and they have high computation demand. A review of clustering validity indexes that are based on external and internal criteria can be found in [6]. Also as was mentioned [7], the validity assessment approaches based on relative criteria work well in non overlapping cases. If the data set considered contains overlapping clusters, then, the majority of the existing validity index fails to detect the right number of clusters.

In this paper, we propose a new and efficient measure to determine the optimal number of clusters, based on Maximum Entropy Principle (MEP), which not only handles efficiently high degree overlapped cases, but also are completely automatic, requiring any parameter determination. We show also in some examples that it works well also in non gaussian mixture models.

The organization of the rest of the paper is as follows. In section 2, we briefly review some validity measures related to our work, and also present some of their shortcomings. In section 3, the proposed measure based on MEP is presented and the correspondent algorithm. Section 4 presents many examples using artificial and real data sets to demonstrate the effectiveness of the proposed measure. Finally, section 5 concludes the paper.

2. Related Work

The Fuzzy C-Means (FCM) clustering algorithm has been widely used to obtain the fuzzy c-partition. This algorithm requires that the user predefine the number of clusters *k*; however, it is not always possible to know the number of clusters in advance. Different fuzzy partitions are obtained at different values of *k*. Thus, an evaluation methodology is required to validate each of the fuzzy c-partitions and, to obtain an optimal partition (or optimal number of clusters c). This quantitative evaluation is the subject of cluster validity. The mathematical formula used to compute the validation is referred to as a cluster validity index.

Many clusters validity indexes for fuzzy clustering are proposed in the literature in order to find an optimal number of clusters. [3] Proposed two cluster validity indexes for fuzzy clustering: Partition Coefficient (V_{PC}) and Partition Entropy (V_{PE}.) These indexes V_{PC} and V_{PE} are sensitive to noise or a weighting exponent m. Other indexes such as V_{FS} and V_{XB} which take into account the geometric properties of input data were proposed respectively [5] and Xie-Beni [15].The V_{FS} index is sensitive to both high and low exponent m. V_{XB} provided a good response over a wide range of choices both for c=2 to 10 and for $1<m \leq 7$. However, V_{XB} decreases monotonically as the number of clusters c becomes very large and close to the number of data n. [12] extended Xie-Beni index V_{XB} to eliminate its monotonic decreasing tendency. To achieve this, a punishing function was introduced to the numerator part of Xie and Beni original validity index. [7] have defined a new validity index V_{S_Dbw}. This latter exploits also the compactness and the separation properties of the data set. The compactness is measured by the cluster variance whereas the separation by the density between clusters. As was mentioned [8] the index V_{S_Dbw} is optimized for data sets that include compact and well-separated clusters that is in non overlapping cases. [10] attempted to determine the optimal number of clusters by measuring the status of the given partition with both an under-partition function and an over-partition function. The proposed index V_{SV} is the sum of the two functions. V_{SV} provides enhanced performances when compared with the previous studies.

More recently, a new validity index V_{OS} was proposed [11], V_{OS} exploits an overlap measure and a separation measure between clusters. The proposed index V_{OS} was defined as the ratio of the overlapping degree to the separation. The overlap measure, which indicates the degree of overlap between fuzzy clusters, is obtained by computing an inter-cluster overlap. The separation measure, which indicates the isolation distance between fuzzy clusters, is obtained by computing a distance between fuzzy clusters. As was mentioned [11] the proposed index V_{OS} is more reliable than other indexes. Unfortunately, from the tests on the IRIS data that have real overlapping clusters, the authors have seen that V_{OS} does not discriminate the two overlapping clusters.

3. The Proposed Validity Index

For a given data set, we obtain, after some clustering process, a partition on k clusters $c_1 \ldots c_j \ldots c_k$. Now, define P_{ij} as a measure of the links between any point i and the cluster c_j, for $j = 1 \ldots k$. As all memberships of any of those clusters c_j are known, we can set $P_{ij}=0$ for $i \notin c_j$ and, for $i \in c_j$, $P_{ij}>0$ are normalized by:

$$\sum_{i \in c_j} P_{ij} = 1 \text{ , for } j = 1 \ldots k. \quad (1)$$

For all the clusters, we have:

$$\sum_{j=1}^{k} \sum_{i \in c_j} P_{ij} = k \quad (2)$$

$$\sum_{j=1}^{k} \sum_{i \in c_j} \left(\frac{P_{ij}}{k} \right) = 1$$

(3)

The entropy of all the clusters is defined by:

$$S = -\sum_{j=1}^{k} \sum_{i \in c_j} \left(\frac{P_{ij}}{k} \right) \ln \left(\frac{P_{ij}}{k} \right)$$

(4)

$$S = -\frac{1}{k} \sum_{j=1}^{k} \sum_{i \in c_j} P_{ij} \ln(P_{ij}) + \ln(k)$$

(5)

$$S = \frac{1}{k} \sum_{j=1}^{k} S_j + \ln(k)$$

(6)

where S_j is given by:

$$S_j = -\sum_{i \in c_j} P_{ij} \ln(P_{ij})$$

(7)

S_j is the entropy corresponding to the cluster j. This entropy will be maximal when all the data points of each cluster have the same association with their cluster centres. Therefore, the optimal number of clusters is the number k whose value of entropy is maximal.

In addition to maximizing the above entropy, we use another constraint which will be minimized. In this second constraint, for each cluster, the nearest neighbour data points to the cluster centre will be privileged. The proposed constraint is given by the following formula:

$$W = \sum_{j=1}^{k} \sum_{i \in c_j} P_{ij} \left\| x_i - g_j \right\|^2$$

(8)

where $\| \ \|^2$ is the euclidean distance, and x_i represents the point i. We are trying to reach the higher possible concentration around or near each cluster centre.

To satisfy the above two constrains, that is to maximize *S* while minimizing *W*, is equivalent to minimize the following expression:

$$T = W - S \qquad (9)$$

$$T = \frac{1}{k} \sum_{j=1}^{k} \sum_{i \in c_j} P_{ij} \, ln\left(P_{ij}\right) - ln(k) + \sum_{j=1}^{k} \sum_{i \in c_j} P_{ij} \left\| x_i - g_j \right|$$

(10)

under *k* constraints: $\sum_{i \in c_j} P_{ij} = 1$ for $j=1,...k$

The lagrange optimizing the formula given in equation 10 under the *k* constraints is given by

$$L = \frac{1}{k} \sum_{j=1}^{k} \sum_{i \in c_j} P_{ij} \, ln(P_{ij}) - ln(k)$$

$$+ \sum_{j=1}^{k} \sum_{i \in c_j} P_{ij} \left\| x_i - g_j \right\|^2 + \sum_{j=1}^{k} \alpha_j \left(\sum_{i \in c_j} P_{ij} - 1 \right)$$

(11)

where α_j is the lagrange multiplicator associated to j^{th} constraint. We then annul the derivation of *L* per P_{ij}:

$$\frac{١}{k} \ln(P_{ij}) + \frac{١}{k} + \left\| x_i - g_j \right\|^{٢} + \alpha_j = ٠$$

(12)

we can then give the expressions of P_{ij} for $i = 1...N$, and $j = 1...k$ by the following:

$$P_{ij} = Z_j^{-1} \exp\left[-k \left\| x_i - g_j \right\|^2 \right]$$

(13)

where Z_j is a normalization coefficient given by:

$Z_j = \exp(1 + k\alpha_j)$

By replacing the expression of P_{ij} given by (13) in the corresponding constraint expression, we obtain the expression of Z_j given below:

$$Z_j = \sum_{i \in c_j} exp\left[-k \left\| x_i - g_j \right\|^{□} \right]$$

(14)

then P_{ij} coefficients can be computed by:

$$P_{ij} = \frac{\exp\left[-k \left\| x_i - g_j \right\|^{٢} \right]}{\sum_{i \in c_j} \exp\left[-k \left\| x_i - g_j \right\|^{٢} \right]}$$

(15)

now, we define, our proposed index V_{MEP} as the whole entropy:

$$V_{MEP} = S = \frac{□}{k} \sum_{j=□}^{k} S_j + ln(k)$$

(16)

where *Sj* is defined by equation 7 which use *Pij* defined in equation 15. The optimal number of clusters is then the number *k* whose value of V_{MEP} is maximal.

The proposed new algorithm using the new index $V_{MEP.}$ We propose in this section our new general algorithm based on the novel index V_{MEP}. The optimal number of clusters is the number *k* whose value of V_{MEP} is maximal. The steps of the algorithm are:

```
A. Fix the maximal number of classes $K_{max}$
B. $K \leftarrow K_{max}$
C. Do while $k \neq 1$,
    C.1.Application of clustering algorithm (exp: call Fuzzy C-means or k-means to define the k classes $c_1 ... c_j ... c_k$. and determine the $g_j$ centres, for $j = 1, ..., k$)
    C.2.Compute the $P_{ij}$ probability with formula 15
    C.3.Compute the $S_k$ entropy with formula 7
    C.4.$K \leftarrow K-1$
    End
D. $V_{MEP} = \max S_k$, for $k = 2, K_{max}$ . ( The correct number of clusters is then the k for which the maximum is due)
```

4. Experimental Results

To test the performance of proposed validity V_{MEP}, we use it to determine the optimal number of clusters in some of synthetic data and also in a well known real data set. However, in earlier publications, V_{SV}, proposed [10], was compared with the following validity indexes V_{PC}, V_{PE}, V_{FS}, V_{XB}, V_K and $V_{S\text{-}Dbw}$. It provides enhanced performances; and in the previous work of one of the authors [13] it was also implemented and used to find the optimal number of clusters using Gaussian Mixture Model (GMM), and the EM algorithm for clustering process, this scheme was successfully applied to extract the design regions in color textile image. Therefore, in this investigation, we will just compare our proposed index V_{MEP} to V_{SV}. In first, we review some applications of V_{MEP} to GMM. We generate sixteen artificial data sets. The first one, DataSet1, is like the well known four polonaise balls [4], Figure 1 shows the scatter plot of this data set; it has 4 compact and well-separated clusters aligned in diagonal. Each cluster was generated using normal distribution; the parameters used for generating this data set are given in Table 1.

Table 1. Parameters used for generating DataSet1.

Cluster Number	Number of Points	Mean Vector	Covariance Matrix
Cluster 1	1000	(-4; -4)	(2 0; 0 2)
Cluster 2	1000	(0; 0)	(1 0; 0 1)
Cluster 3	1000	(4; 4)	(1 0; 0 1)

Cluster 4	1000	(8; 8)	(2 0; 0 2)

The others fifteen Data Sets: DataSet2,…, DataSet15 and DataSet16, are derived from the first one, by producing a two overlapping clusters with different degree of overlap. We move the coordinates centre of cluster 2 having as coordinates centre (0, 0) as shown in Table 1, to a series of the following centers coordinates (1, 1), (1.5; 1.5), (1.6; 1.6), (1.7; 1.7), (1.8; 1.8), (2; 2), (2.5; 2.5), (2.9; 2.9), (3; 3), (3.25; 3.25), (3.5; 3.5), (3.6; 3.6), (3.7; 3.7), (3.9; 3.9), and finally (4; 4) which are the coordinates centre of cluster 3, as shown in Table 1.

We apply V_{SV} and V_{MEP} to these data sets, and we see if V_{MEP} can perform V_{SV}? If yes, how well does it, and up what limit?. The cluster validation results using V_{SV} and V_{MEP} are shown in Figure 1. For the DataSet1, having well-separated clusters, both V_{SV} and V_{MEP} can select correctly 4 as optimal number of clusters. For the DataSet2 to DataSet4, which have two overlapping clusters with low degree of overlap, also both V_{SV} and V_{MEP} select correctly 4 as the optimal number of clusters. For DataSet5, V_{SV} select 3 which is a failure result. By increasing the degree of overlap in DataSet6, DataSet7, V_{SV} also fails, it select 3 which is not a correct optimal number of clusters. Instead, V_{MEP} selects correctly 4 clusters for all these data sets (DataSet5, DataSet6, and DataSet7).

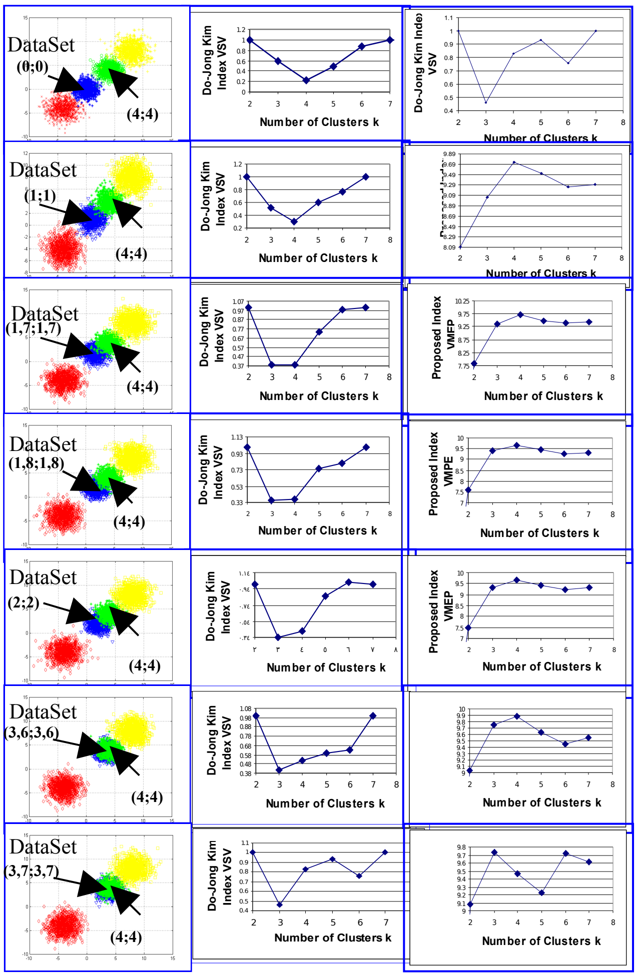

Figure 1. Results of clusters validation using V_{SV} (minimal value) and V_{MEP} (maximal value), displayed from DataSet1 to DataSet7.

From the above results, we conclude that V_{SV} can work correctly only in the presence of a low degree of overlap, and it produces a failure results when dealing with relatively a high overlapping degree. We then stop to apply V_{SV} to data sets having a superior overlapping degree such as DataSet8…DataSet16; and we continue to apply only V_{MEP}.

The result of applying V_{MEP} to the DataSet8 to DataSet13, are presented respectively in Figure 2. V_{MEP} can still work well; it selects correctly 4 as the optimal number of clusters. In DataSet14…DataSet16, the centre coordinates of the moved cluster number 2 are respectively (3.7; 3.7), (3.9; 3.9), and (4; 4). These centers are very close to those of the fixed cluster number 3 whose coordinates centre are (4; 4). This yields a very high overlapping degree. In this case, we can see in Figure 3 that the two overlapping clusters represent approximately one cluster. V_{MEP} can not select 4 as optimal number of clusters, it select 3 clusters which can be considered as evident result.

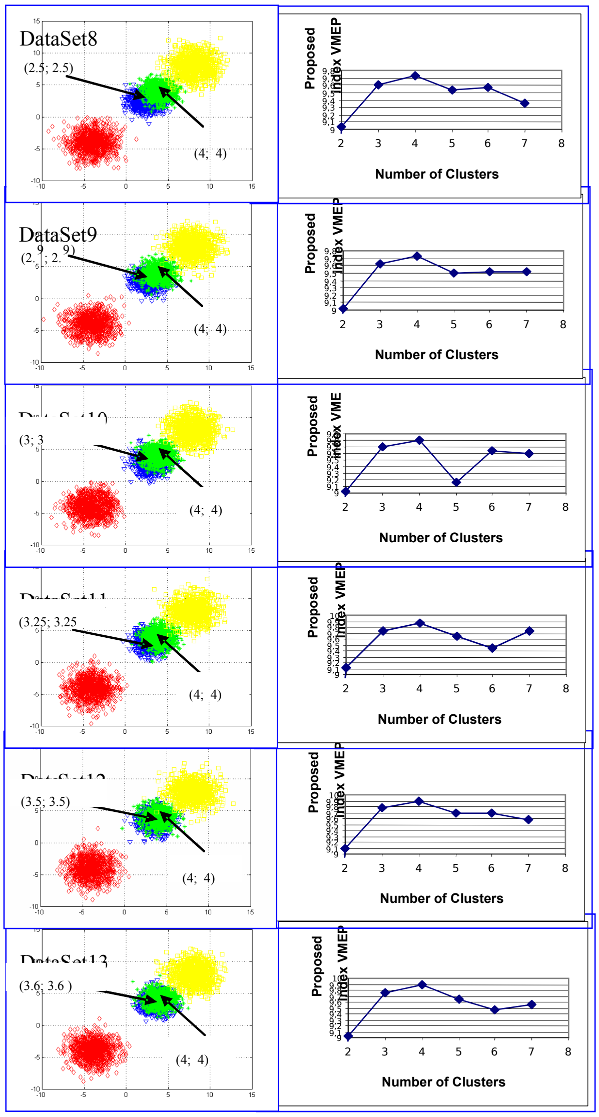

Figure 2. Results of clusters validation using the proposed V$_{MEP}$, displayed from DataSet8 to DataSet13

Then, we see that V_{MEP} performs clearly V_{SV}, it can still work well and select the correct optimal number of clusters for all data sets up DataSet13, as shown in Figures 2 and 3. From the data sets DataSet14 up DataSet16, as shown in Figure 3, V_{MEP} can not select 4 as optimal number of clusters; because the clusters number 2 and 3 are extremely overlapped, and they may be regarded as one cluster. We conclude that V_{SV} can work correctly only in the presence of a low degree of overlap, and it produces a failure results when dealing with relatively high overlapping degree, while V_{MEP} still works well, and selects correctly 4 as optimal number of clusters with very high overlap. We conclude that V_{MEP} performs clearly V_{SV} for GMM as verified in our early work [1].

The performance of V_{MEP} is also examined using a well known real Iris data set [2]. Results are shown in Figure 4. Both the two index V_{MEP} and V_{SV} select correctly 3 as optimal number of clusters. Here, V_{SV} can work well because the low degree of overlap.

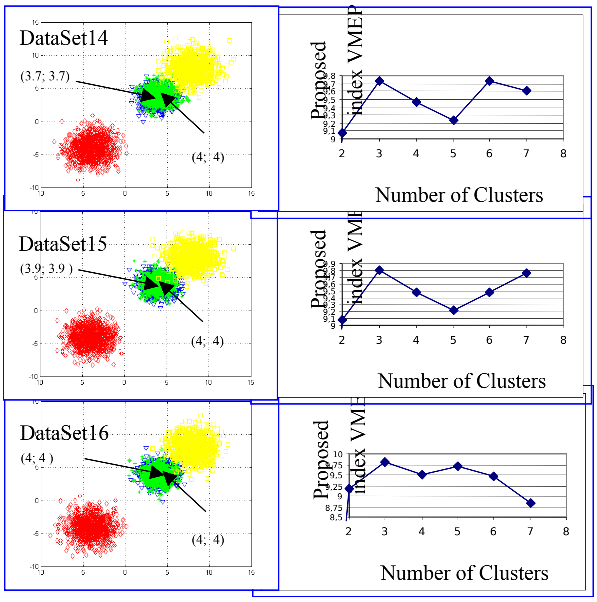

Figure 3. Results of clusters validation using the proposed V$_{MEP}$, displayed from DataSet14 to DataSet16

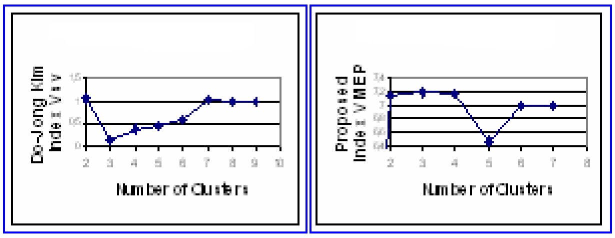

Figure 4. Results of clusters validation using Do-Jong Kim's index V$_{SV}$ (minimal value), and the proposed V_{MEP} (maximal value), applied to the Iris Data Set.

Now, what about non GMM, Figure 5 shows results when V_{MEP} is applied to other forms like banana forms. In the present work, we generate 4 banana forms named respectively Banana set1, Banana set2, Banana set3, Banana set4. In all of them, V_{MEP} detects the correct and real number of clusters.

Banana set1 describe two banana forms enclosed into one circle which is wrapped by one banana form. The result of applying V_{MEP} to the Banana set1 shows that it can select 4 clusters which is the correct number of clusters for banana set1. For banana set2, we stay the same two banana forms enclosed now in two symmetric banana forms with same centre but with different radius. In this case V_{MEP} can select also 4 clusters which is the correct number of clusters for banana set2. The illustration of the banana set3 show two symmetric banana forms with same centre and same radius. We keep into them the same two banana forms enclosed in banana set1 and banana set2. V_{MEP} works also well and selects 3 clusters which is the logic and correct number of clusters. Finally, we test our new index on a combination of different forms and overlapping case. The result of this application showing in graphic BSet4 is very interesting. V_{MEP} can detect 5 clusters which is the correct number of clusters. This last result completes the performance and the robustness of $V_{MEP.}$

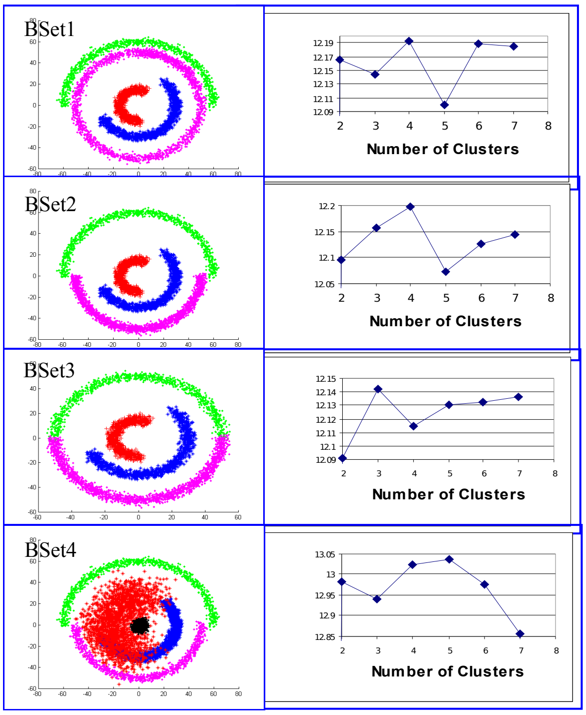

Figure 5. Results of clusters validation using V$_{MEP}$ for some banana forms.

5. Conclusion

We introduced in this paper a new formulation of a cluster validity index for the validation of the fuzzy *k*-partitions that are generated by the application of the FCM clustering algorithm. This new index can be playing an important role in solving many problems in pattern recognition, amelioration of the quality of products in marketing. The proposed index V_{MEP} is based on the MEP and the optimal number of clusters is the number k^* whose value of V_{MEP} is maximal. The performance of our index V_{MEP} was examined, in both our generated synthetic data sets and in real data example and a robustness of this new index is completed by the extension of the method to non-GMM with overlap. The experimental results show the superiority of our measure V_{MEP} to the existing ones and its capacity to detect the "right" number of clusters with different shapes and degree of overlap.

Finally, we report also another advantage of our index. The definition of V_{MEP} uses any parameter produced by the adopted clustering algorithm. Therefore, V_{MEP} is independent of any clustering algorithm. This allows us to choose any one, such as Gustafson-Kessel (GK) algorithm which can deals with ellipsoidal clusters, or EM clustering algorithm. This will be the subject of our next investigation.

References

[1] Ammor O., Lachkar A., Slaoui K., and Rais N., "New Efficient Approach to Determine the Optimal Number of Clusters in Overlapping Cases," *in Proceedings of the IEEE on Advances in Cybernetic Systems*, UK, pp. 26-31, 2006.

[2] Anderson E., "The IRISes of the Gaspe Peninsula," *Bulletin of the American Iris Society*, vol. 59, no. 1935, pp. 2-5, 1959.

[3] Bezdek J., "Cluster Validity with Fuzzy Sets," *Cybernetics and Systems*, vol. 3, no. 3, pp. 58-72, 1975.

[4] Cembrzynski T., "Banc D'essai Sur Les Boules Polonaises, Des Trois Critères de Décision Utilisés Dans La Procédure de Classification, MNDOPT Pour Choisir un Nombre de Classes," *RR-0784 Rapport de recherche de l'INRIA,* 1986.

[5] Fukuyama Y. and Sugeno M., "A New Method of Choosing the Number of Clusters for the Fuzzy C-Means Method," *in Proceedings of the 5th Fuzzy Systems Symposium*, Japan, pp. 247-250, 1989.

[6] Halkidi M., Batistakis Y., and Vazirgiannis M., "Quality Scheme Assessment in the Clustering Process," *in Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, London, pp. 265-276, 2000.

[7] Halkidi M. and Vazirgiannis M., "Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set," *in Proceedings of 1st IEEE International Conference on Data Mining (ICDM'2001)*, USA, pp. 187-194, 2001.

[8] Halkidi M., Batistakis Y., and Vazirgiannis M., "Clustering Validity Checking Methods: Part II," *Special Interest Group on Management of Data (SIGMOD)*, vol. 31, no. 3, pp. 19-27, 2002.

[9] Jain A., Murty M., and Flynn P., "Data Clustering a Review," *ACM Computing Surveys*, vol. 31, no. 3 , pp. 264-323, 1999.

[10] Kim D., Park Y., and Park D., "A Novel Validity Index for Determination of the Optimal Number of Clusters," *IEICE Transactions on Information System*, vol. D-E84, no. 2, pp. 281-285, 2001.

[11] Kim D., Kwang A, Lee H., and Lee D., "On Cluster Validity Index for Estimation of the Optimal Number of Fuzzy Clusters," *Pattern Recognition*, vol. 37, pp. 2009-2025, 2004.

[12] Kwon S., "Cluster Validity Index for Fuzzy Clustering," *Pattern Recognition Letters*, vol. 34, no. 22, pp. 2176-2177, 1998.

[13] Lachkar A., Benslimane R., D'Orazio L., and Martuscelli E., "A System for Textile Design Patterns Retrieval Part 1: Design Patterns Extraction by Adaptive and Efficient Color Image Segmentation Method," *Journal of the Textile Institute*, no. 10.1533. joti, 124r1, 2005.

[14] Liu J. and Yang Y., "Multiresolution Color Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 7, pp. 689-700, 1994.

[15] Liu X. and Beni G., "A Validity Measure for Fuzzy Clustering," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 13, no. 8, pp. 841-847, 1991.

[16] Sharma S., *Applied Multivariate Techniques*, John Wiley and Sons, 1996.

[17] Trivedi M. and Bezdek J., "Low-Level Segmentation of Aerial Images with Fuzzy Clustering," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 16, no. 4, pp. 589-598, 1986.

Ouafae Ammor is a professor in the Department of Mathematics at Faculty of Sciences and Technology of Fes, Morocco. She has obteined her Master's degree in statistics from Polytechnique school of Montreal, Canada. Her research focuses on the clustering in data analysis, pattern recognition, colour image segmentation,

medical image processing, fuzzy logic and spatial data analysis.

Abdelmonaime Lachkar received the Master degree in automatic and systems analysis in 1999, and his PhD degree from the USMBA, Morocco in 2004. In computer sciences. He is associate professor in Computer Sciences in ESTM at Moulay Ismail University-Morocco. His current research interests include image indexing and retrieval, shape representation, indexing and retrieval in large shapes databases, colour image segmentation, unsupervised clustering, cluster validity index, pattern recognition, arabic and latin handwritten recognition, document clustering and categorisation, medical image processing, image compression and watermarking.

Khadija Slaoui is a professor in the Department of Physics at Faculty of Sciences Dhar Mehraz of Fes, Morocco. She has obteined her Master's in image processing at Polytechnique Institute of Toulouse, France, and PHD in data analysis at University Sidi Mohammed Ben Abdellah. Her research focuses on signal processing, clustering, and images processing.

Noureddine Rais is full professor at the University of Fez in Morocco. He has a PhD in mathematical statistics from the University of Montreal, Canada and a 3rd Cycle Doctorate in Statistics and Mathematical Models from University of Paris-sud, Orsay, France. His current research interests covers bootstrap, spatial data, survey sampling, decision theory, data analysis, data mining, experimental design, linear models and analysis of variance, with applications in many fields, specially statistical processing control, arabic handwritten recognition, and medical image processing.