IADIS International Journal on Computer Science and Information Systems
Vol. 4, No. 1, pp. 75-84
ISSN: 1646-3692

# Assessing the Quality of Fuzzy Partitions in Overlapping Data Sets Using Maximum Entropy Principle

*A. O. Ammor[1], A. Lachkar[2], K. Slaoui[3] and N. Rais[4]*

*[1] Laboratoire de Modélisation et Calcul Scientifique. FSTF Université Sidi Mohammed Ben AbdellahFès, Morocco. E-mail : w_ammor@yahoo.fr.*

*[2] Laboratoire d'Electronique, Informatique et de Biotechnologie LEIB. Dept Informatique. E.S.T.M, Université Moulay Ismail. Morocco.*

*[3] Laboratoire LESSI .Faculté des sciences Dhar Mehraz, Université Sidi Mohammed Ben Abdellah, Fès, Morocco.*

*[4] Laboratoire ISQ. Faculté des Sciences Dhar Mehraz, Université Sidi Mohammed Ben Abdellah Fès, Morocco*

**Abstract**

Many validity indexes have been proposed for evaluating clustering results. They usually have a tendency to fail in selecting the right number of clusters when dealing with overlapping clusters such as the IRIS data. To overcome this limitation, we propose in this paper, a new cluster validity index based on Maximum Entropy Principle, named $V_{MEP}$. $V_{MEP}$ allows finding the correct number of clusters, and can deal successfully with or without the presence of overlap, even when this later is higher between clusters. Many simulated and real examples are presented, showing the superiority of $V_{MEP}$ to the existing indexes.

## 1. Introduction

Fuzzy c-means FCM clustering algorithms has been widely used to obtain fuzzy c-partition. This algorithm requires a fixed number of clusters k. Different fuzzy partitions are obtained for 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 k*. Finding the "right" number of clusters, k*, for a data set, is a difficult and often ill-posed problem. We introduce hereafter a new cluster validity index, $V_{MEP}$, based on Maximum Entropy Principle and a measure of cluster compactness. $V_{MEP}$ is parameter free, works well and detects the correct number of clusters even when dealing with high overlapping clusters. We present in this paper some results on simulated and real examples which illustrate the superiority of $V_{MEP}$ to the existing indexes. They show also that our new index is performing not only for Gaussian models but also with different shapes of clusters with or without overlap.

## 2. Related work

Many clusters validity indexes for fuzzy clustering are proposed in the literature [1-4] in order to find an optimal number of clusters. Bezdek [5] proposed: Partition Coefficient $V_{PC}$ and Partition Entropy $V_{PE}$. These indexes are sensitive to noise or a weighting exponent m. $V_{FS}$ and $V_{XB}$ are proposed respectively by Fukayama and Sugeno [6] and Xie-Beni [7]. 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 k=2 to 10 and for $1<m \le 7$. However, $V_{XB}$ decreases monotonically as the number of clusters k becomes very large and close to the number of data n. Kwon et al. introduced a punishing function to the numerator part of $V_{XB}$ to eliminate its monotonic decreasing [8]. Maria Halkidi [9] defined a $V_{S\_Dbw}$ which performs well when clusters are compacts and well separated, i.e. in the non overlapping clusters cases. In 2001, Do-Jong Kim [10] proposed index $V_{SV}$ which provides enhanced performances when compared with the previous studies.

As seen, there are no many indexes for the overlapping cases. One of the most recent is $V_{OS}$, proposed by Dae-Won Kim et al. in 2004 [11]. $V_{OS}$ is defined as the ratio of an overlap and a separation measures between clusters. As was mentioned by the authors [11], the proposed index $V_{OS}$ is more reliable than other indexes. Unfortunately, from the tests on the IRIS data, which 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 $c_1 \dots c_j \dots c_k$. Now, define $P_{ij}$ as a measure of the links between any point i and the cluster $c_j$, for j = 1…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:

$$\sum_{i \in c_j} P_{ij} = 1, \text{ for } j = 1 \dots k \qquad (1)$$

For all the clusters, we have:

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

$$\sum_{j=1}^{k} \sum_{i \in c_j} \left( \frac{P_{ij}}{k} \right) = 1 \qquad (3)$$

The entropy of all the clusters is defined by:

$$S = -\frac{1}{k} \sum_{j=1}^{k} \sum_{i \in c_j} P_{ij} \ln\left(P_{ij}\right) + \ln(k) \qquad (4)$$

$$S = \frac{1}{k} \sum_{j=1}^{k} S_j + \ln(k) \qquad (5)$$

Where $S_j$ is given by:

$$S_j = -\sum_{i \in c_j} P_{ij} \ln\left(P_{ij}\right) \qquad (6)$$

$S_j$ 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 privilege nearest neighbor data points to the cluster centre, we shall also minimize a second constraint:

$$W = \sum_{j=1}^{k} \sum_{i \in c_j} P_{ij} \left\| x_i - g_j \right\|^2 \qquad (7)$$

where $\| \ \|^2$ is the Euclidean distance, $x_i$ represents the point i and $g_j$ the centre of cluster $c_j$. 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:

$$T=W-S \tag{8}$$

$$T=\frac{1}{k}\sum_{j=1}^{k}\sum_{i\in c_j}P_{ij}\ln\left(P_{ij}\right)-\ln(k)+\sum_{j=1}^{k}\sum_{i\in c_j}P_{ij}\left\|x_i-g_j\right\|^2 \tag{9}$$

This minimization must be done under the k constraints in (1):

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

The Lagrange obtained is given by:

$$L=\frac{1}{k}\sum_{j=1}^{k}\sum_{i\in c_j}P_{ij}\ln(P_{ij})-\ln(k)+\sum_{j=1}^{k}\sum_{i\in c_j}P_{ij}\left\|x_i-g_j\right\|^2+\sum_{j=1}^{k}\alpha_j\left(\sum_{i\in c_j}P_{ij}-1\right) \tag{10}$$

Where $\alpha_j$ is the Lagrange multiplicator associated to $j^{th}$ constraint. We then annul the derivation of L per $P_{ij}$:

$$\frac{1}{k}\ln(P_{ij})-\frac{1}{k}+\left\|x_i-g_j\right\|^2+\alpha_j=0 \tag{11}$$

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

$$P_{ij}=Z_j^{-1}\exp\left[-k\left\|x_i-g_j\right\|^2\right] \tag{12}$$

where $Z_j$ is a normalization coefficient given by:

$$Z_j=\exp\left(1+k.\alpha_j\right)$$

By replacing the expression of Pij given by (13) in the corresponding constraint expression, we obtain the expression of $Z_j$ given below:

$$Z_j=\sum_{i\in c_j}\exp\left[-k\left\|x_i-g_j\right\|^2\right] \tag{13}$$

Then Pij coefficients can be computed by :

$$P_{ij}=\frac{\exp\left[-k\left\|x_i-g_j\right\|^2\right]}{\sum_{i\in c_j}\exp\left[-k\left\|x_i-g_j\right\|^2\right]} \tag{14}$$

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

$$V_{MEP}=S=\frac{1}{k}\sum_{j=1}^{k}S_j+\ln(k) \tag{15}$$

where $S_j$ is defined by (6) which use $P_{ij}$ defined in equation (14). The optimal number of clusters is then the number $k^*$ whose value of $V_{MEP}$ is maximal.

## 4. Experimentals results

The $V_{SV}$, proposed by Do-Jong Kim et al in 2001 [10], was compared in earlier publications with the following validity indexes $V_{PC}$, $V_{PE}$, $V_{FS}$, $V_{XB}$, $V_K$ and $V_{crit}$. This validity index $V_{SV}$ provides enhanced performances.

To test the performance of the proposed validity $V_{MEP}$, we use it to determine the optimal cluster numbers in some of synthetic data and also in a well known real data set.

We generate sixteen artificial data sets. The first one, DataSet1, is like the well known Four Polonaise Balls [12]. 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 with parameters given in table-1 below:

| 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) |

**Table- 1:** parameters used for generating DataSet1

The others fifteen Data Sets, named: DataSet2 … DataSet16, are derived from the first one, DataSet1, by moving the centre (0, 0) of cluster 2 in direction of the centre (4, 4) of cluster 3, producing hence two overlapping clusters. Coordinates of new centers of the cluster 2 are (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 (table-1). Figure-1, figure-2¸ and figure-3 show the generated data sets with two overlapping clusters (clusters 2 and 3) with increasing degree of overlap.

Now, we apply $V_{SV}$ and $V_{MEP}$ to these Data Sets, and we will see if our proposed clusters validity index $V_{MEP}$ can performs $V_{SV}$? 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, DataSet3 and, 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).

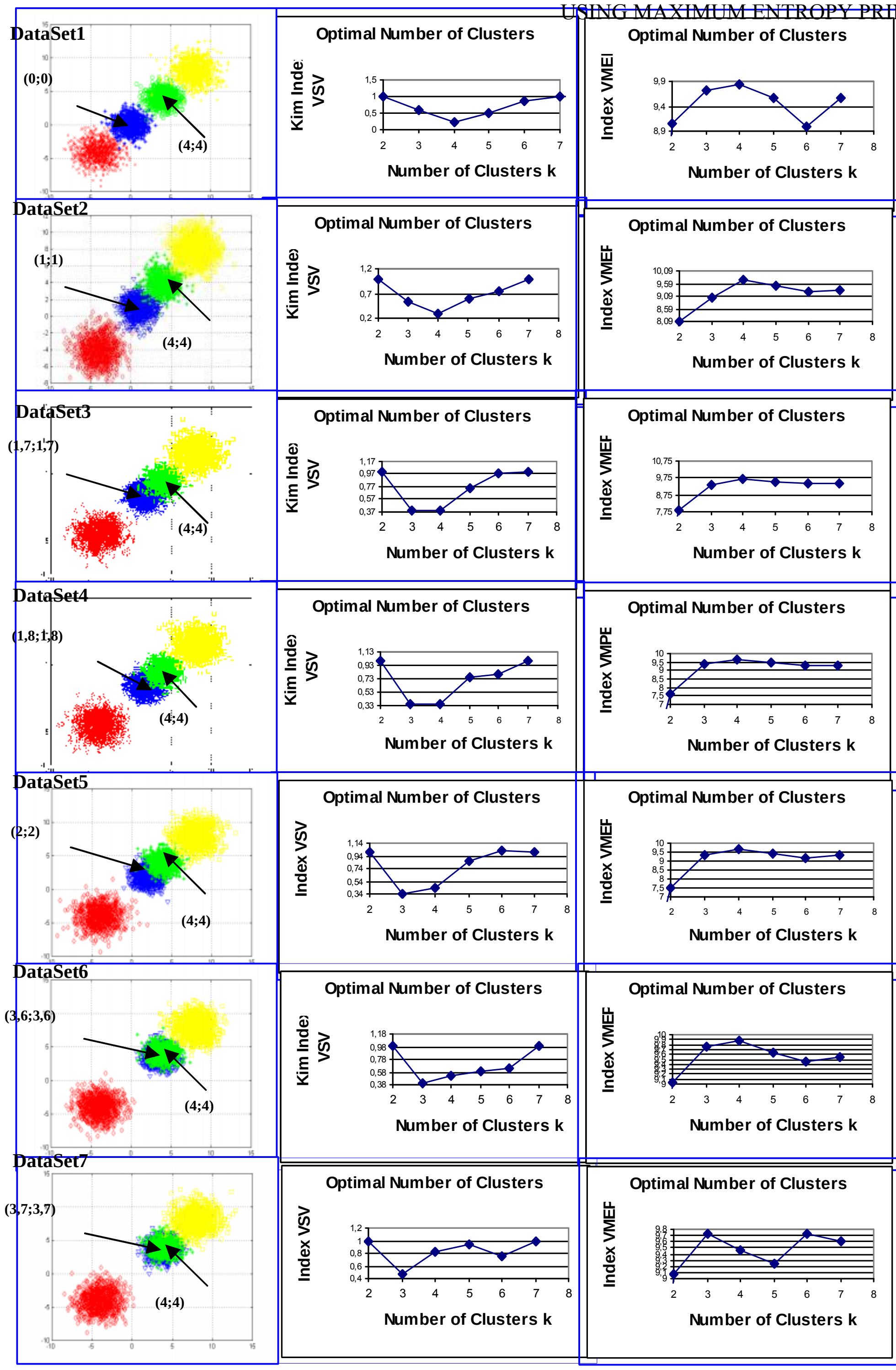


**Figure-1:** Results of clusters validation using Do-Jong Kim's index $V_{SV}$ (minimal value), and the proposed $V_{MEP}$ (maximal value), displayed from DataSet1 to DataSet7

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 result when dealing with relatively 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…DataSet13, are presented respectively in figure-2, these latter show that $V_{MEP}$ 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 –which overlap with the fixed cluster number 3- 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 selects 3 clusters, which can be considered as evident and logical result.

We conclude that the proposed validity index $V_{MEP}$ performs clearly $V_{SV}$, it can still select the correct optimal number of clusters for the data sets DataSet5, DataSet6, and

DataSet7 (figure-1), for which $V_{SV}$ gives a failure result. And also, for DataSet8 up DataSet13 (figure-2), $V_{MEP}$ can still work well.

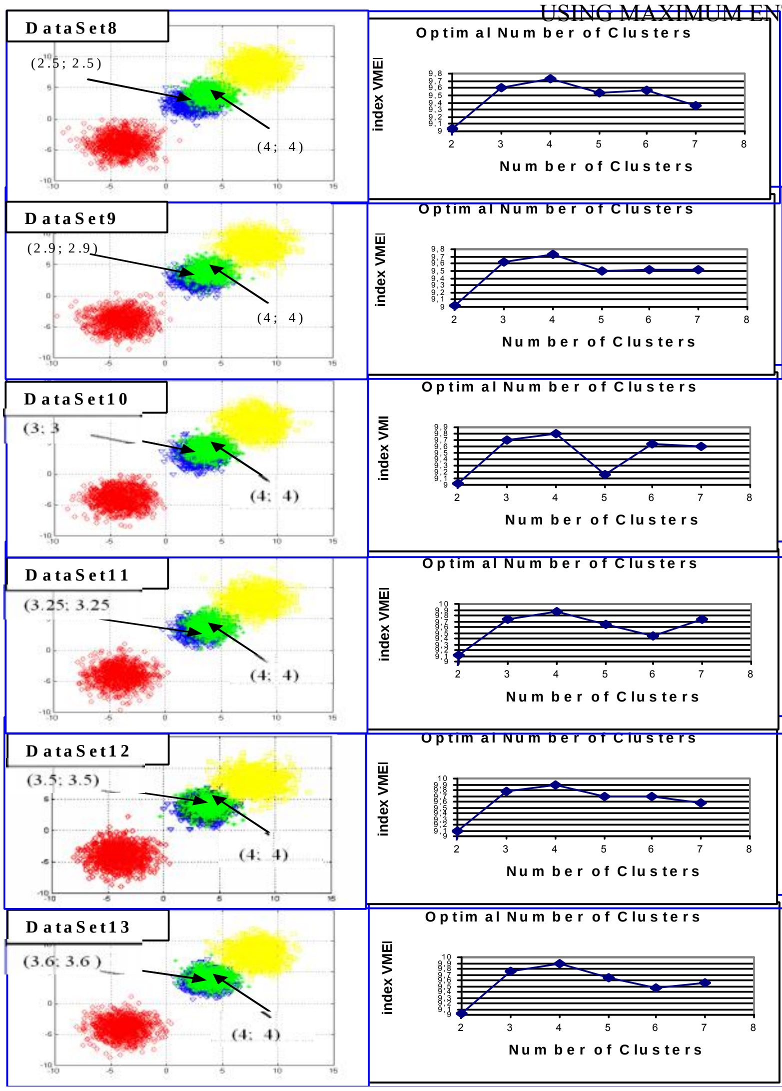


**Figure-2.** Results of clusters validation using the proposed $V_{MEP}$, displayed from DataSet8 to DataSet13

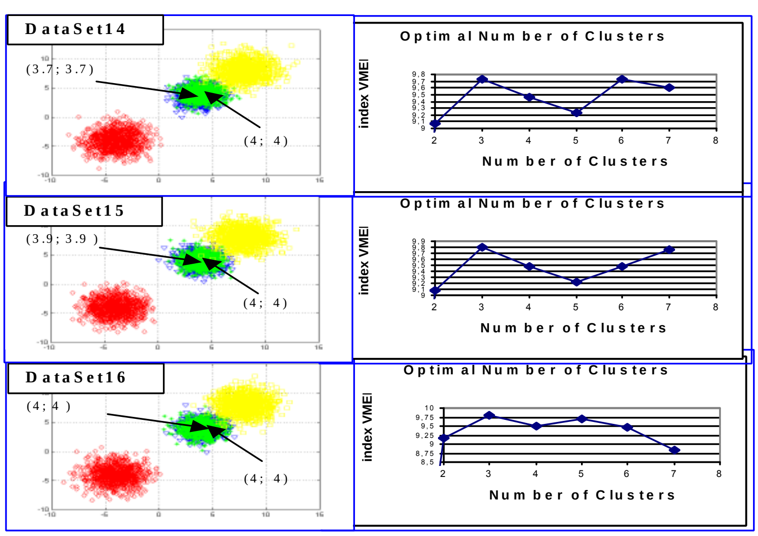


**Figure-3.** Results of clusters validation using the proposed $V_{MEP}$, displayed from DataSet14 to DataSet16

The performance of the proposed index $V_{MEP}$ is also examined using the well known real Iris Data Set [13]. It consists of 150 biometric measurements in the four-dimensional space. Iris Data are grouped into 3 clusters of 50 data points each, namely: Setosa, Versicolor, and Virginica. Most of the recent indexes presented in the literature fails to handle with the Iris data sets.

More recently, in 2004, Dae-Won Kim et al [11] proposed a new index named $V_{OS}$ that uses the concept of the degree of Overlap and Separation. $V_{OS}$ was developed specially for handling with overlapping cases. As was mentioned by the authors [11], when applied to the Iris Data Sets, $V_{OS}$ was unable to detect the correct number of cluster 3. It selects 2 as optimal number of clusters, which is a failure result.

In figure-4, we present the results of applying $V_{SV}$ and $V_{MEP}$. Both select correctly 3 as optimal number of clusters. Here $V_{SV}$ can work well because the low degree of overlap.

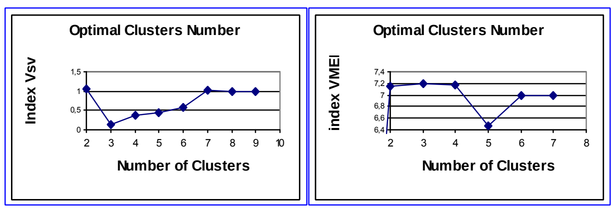


**Figure-4**: Results of clusters validation using Do-Jong Kim's index $V_{SV}$ (minimal value), and the proposed $V_{MEP}$ (maximal value), applied to the Iris Data Set.

We conclude that the proposed validity index $V_{MEP}$ performs clearly $V_{SV}$ at least for Gaussian mixtures models as verified in our early work [14].

Now, what about non Gaussian mixtures models? Fig 5 shows results when $V_{MEP}$ is applied to banana forms. In the present work, we generate 4 banana forms named respectively BSet1, BSet2, BSet3, BSet4. In all of them, $V_{MEP}$ detects the correct and real number of clusters.

BSet1 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.

For BSet2, 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.

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 is very interesting. $V_{MEP}$ can detect 5 clusters which is the correct number of clusters. This last result illustrates the performance and the robustness of $V_{MEP}$.

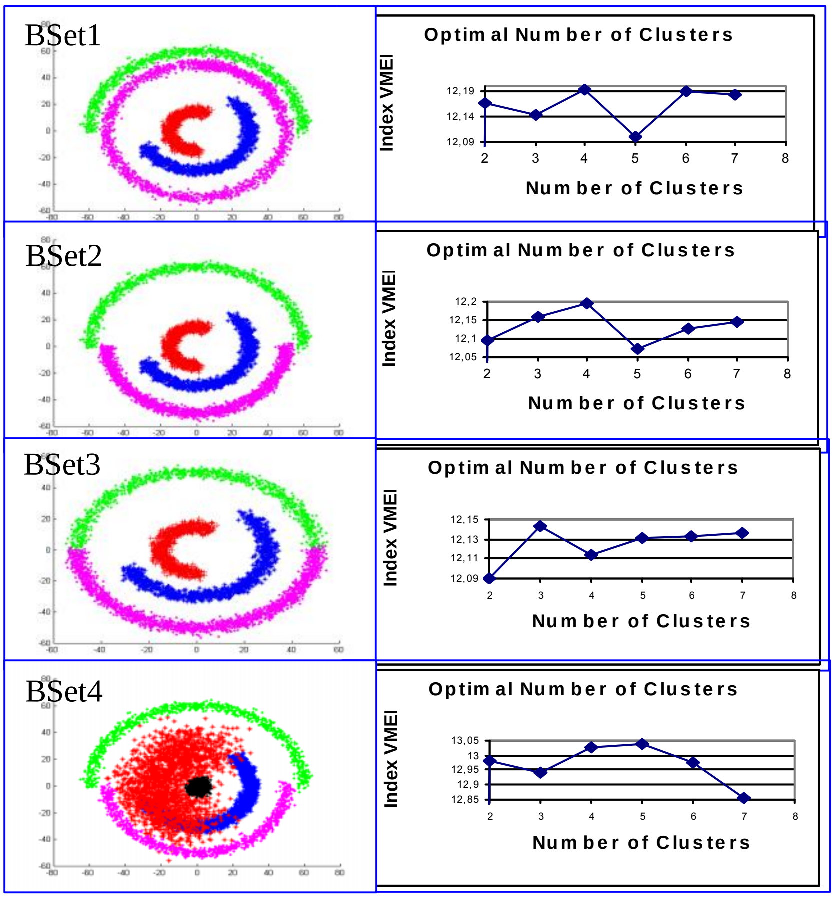


**Figure 5***:* Results of clusters validation using $V_{MEP}$ for some banana forms.

## 5. Conclusion

A new index is proposed for the validation of the fuzzy c-partitions that are generated by the application of the fuzzy c-means clustering method. The proposed index $V_{MEP}$ is based on the Maximum Entropy Principle. The optimal number of clusters is then 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 another advantage when it can detect the correct number of clusters for not only Gaussian models but also for other shapes.

The experimental results show the superiority of our measure $V_{MEP}$ to the existing ones. Therefore, the proposed clusters validity index $V_{MEP}$ can be used as a reliable tool to evaluate the partitions produced by the application of the fuzzy c-means clustering algorithm. The robustness of our new index is showed also with variety banana forms. $V_{MEP}$ work well not only in this case but can detect a correct and optimal number of clusters when we combine different banana forms with overlapping case.

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 deal with ellipsoidal clusters, or EM clustering algorithm. This will be the subject of our next investigation.

# Reference

[1] A. K. Jain, M. N. Murty and P. J. Flynn: Data clustering: a review, ACM Computing Surveys, Vol. 31, No. 3, pp. 264-323, 1999

[2] S. Theodoridis and K. Koutroubas: Pattern Recognition, Academic Press, 1999

[3] J. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum, 1981.

[4] J. Hartigan, Clustering Algorithms. New York: Wiley, 1975.

[5] J. C., Bezdek, 1974. “Cluster validity with fuzzy sets”, J. Cybernit.3, 58 –72.

[6] Y. Fukuyama, M. Sugeno, “A new methodof choosing the number of clusters for the fuzzy c-means method”, Proceedings of the Fifth Fuzzy Systems Symposium,1989, pp. 247 –250.

[7] L.Xie,G.Beni, “A validity measure for fuzzy clustering”, IEEE Trans.Pattern Anal.Mach.Intell.13(8) (1991) 841–847.

[8] S.H.Kwon,, Cluster validity index for fuzzy clustering, Electron.Lett.34(22) (1998) 2176–2177.

[9] M. Halkidi and M. Vazirgiannis “Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set”, Proc. of ICDM 2001, pp. 187-194, 2001

[10] D. J. Kim, Y. W. Park, and D. J. Park, “A novel validity index for determination of the optimal number of clusters”, IEICE Trans. Inform.Syst.D-E84 (2)(2001)281 –285.

[11] D. W. Kim, A. Kwang, H. Lee and D. Lee, “On cluster validity index for estimation of the optimal number of fuzzy clusters”, Pattern Recognition. Vol 37, pp. 2009 –2025, 2004.

[12] Cembrzynski, T. Banc d'essai sur "les boules polonaises", des trois criteres de decision utilises dans la procedure de classification MNDOPT pour choisir un nombre de classes. RR-0784 Rapport de recherche de l'INRIA.

[13] Anderson E. The IRISes of the Gaspe peninsula. Bull Am IRIS Soc 1935;59:2–5.

[14] O. Ammor, A. Lachkar, K. Slaoui and N. Rais, “New Efficient Approach to Determine the Optimal Number of Clusters in Overlapping Cases”, proceeding of the IEEE on Advances in Cybernetic Systems, pp: 26-31, 2006.