Détermination du nombre optimal de classes présentant un fort degré de chevauchement

Ammor.O*- Raiss.N- Slaoui.K*****

**Laboratoire de modélisation et calcul scientifique. Faculté des sciences et techniques Fès Université Sidi Mohammed Ben Abdellah*

***Laboratoire ISQ. Faculté des sciences Dhar Mehraz, Université Sidi Mohammed Ben Abdellah , Fès, Maroc*

****Laboratoire LESSI . Faculté des sciences Dhar Mehraz, Université Sidi Mohammed Ben Abdellah , Fès, Maroc*

Adresse de correspondance : w_ammor@yahoo.fr

***RÉSUMÉ* :** *Dans cet article, nous présentons un nouvel indice pour la détermination du nombre optimal et correct de classes nommé V_{MEP} basé sur le Principe du Maximum d'Entropie. Les performances de ce nouvel indice déduit d'une combinaison originale entre des méthodes d'analyse des données et le critère du maximum d'entropie, sont montrées à travers un ensemble d'exemples simulés et réels. La procédure est complètement automatique dans le sens qu'elle ne nécessite aucun paramètre de réglage. V_{MEP} montre une grande robustesse, et une supériorité par rapport à d'autres indices déjà existants et assez récents, particulièrement dans le cas du chevauchement spatial entre classes.*

***ABSTRACT**: In this paper, we propose a new and efficient clusters validity measure named V_{MEP} for determination of the optimal and correct number of clusters based on the maximum entropy principle. The performance of this new index which has been shown in by many simulated and real examples is deducted from original combination of data analysis methods and the maximum entropy principle criterion. The method does not require any parameter adjustment, it is then completely automatic. Our new index V_{MEP} shows high robustness and superiority to the existing and recent ones, especially in overlapping clusters case.*

MOTS – CLÉS : Classification non supervisée, Principe du Maximum d'Entropie, chevauchement de classes, nombre optimal de clusters.

KEY WORDS: unsupervised classification, the maximum entropy principle, overlapping clusters, optimal number of clusters.

1. Introduction :

La classification est une notion qui intervient fréquemment dans la vie courante. En effet, il est souhaitable de regrouper les éléments d'un ensemble hétérogène, en un nombre restreint de classes les plus homogènes possibles. Son application a joué un rôle très important pour résoudre plusieurs problèmes en reconnaissance des formes, imagerie, segmentation d'images couleur, data mining et dans différents domaines comme la médecine, la biologie, le marketing,…etc.

Nous parlons de classification non supervisée, ou regroupement, lorsqu'on ne dispose d'aucune information a priori sur les variables à traiter ; et de classification supervisée autrement. Le travail développé dans cette recherche s'inscrit dans le cadre des techniques de classification non supervisée, qui s'apparente à la recherche des groupes homogènes au sein d'un mélange multidimensionnel où le nombre de groupes est inconnu. Les résultats de classification obtenus dépendent fortement du nombre de classes fixé. Il est donc primordial de choisir le nombre exact de classes pour espérer avoir une bonne qualité de classification. Ceci n'est pas toujours simple, surtout en présence de cas de chevauchement entre clusters.

Plusieurs approches ont été proposées sur ce sujet dans différentes applications [6], [11], [12], [13]. Cependant, pour les mêmes données, on peut obtenir des résultats différents selon le nombre de classes k fixé par l'utilisateur. Pour des classes bien séparées, les algorithmes de classification retrouvent généralement le même nombre de clusters. Le problème se pose dans le cas de chevauchement de classes : rares sont les algorithmes qui arrivent à détecter le nombre réel de classes, et ils deviennent invalides pour un degré de chevauchement relativement fort.

Le processus d'évaluation des résultats des algorithmes de classification est appelé indice de validité des clusters. Trois critères sont en général utilisés [8]: Externe, Interne et Relatif. Les deux premiers sont basés sur des méthodes statistiques et demandent beaucoup de temps de calcul [9]. Les techniques basées sur le Critère Relatif mentionné par Maria et al [10] fonctionnent correctement dans le cas de classes compactes et sans chevauchement. Cependant, plusieurs applications présentent différents degrés de chevauchement, et l'application de ces algorithmes reste limitée. Pour surmonter cette limitation, nous proposons dans cet article un nouvel indice de validité pour la détermination du nombre optimal de classes particulièrement celles présentant un fort degré de chevauchement.

Dans la prochaine section, nous présentons quelques critères de validité les plus utilisés, ainsi que leurs limites et inconvénients. La section 3 détaillera notre nouvel indice de validité nommé V_{MEP}. Les résultats expérimentaux sur des exemples réels et artificiels sont présentés dans la section 4, montrant l'efficacité et la robustesse de notre nouvel indice. On finira par la conclusion dans la section 5.

2. Quelques indices de validité existants

Les algorithmes de classification floue (Fuzzy C-means FCM) ont été largement utilisés pour obtenir les k-partitions floues. Cet algorithme suppose la fixation a priori du nombre de classes k par l'utilisateur, ce qui n'est pas toujours possible. Différentes partitions sont ainsi obtenues pour différentes valeurs de k. Une méthodologie d'évaluation est requise pour déterminer le nombre optimal de clusters k^*. C'est ce qu'on appellera indice de validité des clusters (cluster validity index).

Le processus pour le calcul de l'indice de validation des clusters est résumé par les étapes suivantes:

Etape 1 : Initialiser les paramètres des FCM excepté le nombre de clusters k.
Etape 2 : Appliquer l'algorithme FCM pour différentes valeurs de k avec k=2,3,…,cmax. (cmax est fixé par l'utilisateur).
Etape 3 : Calculer l'indice de validité pour chaque partition obtenue à l'étape 2.
Etape 4 : Choisir le nombre optimal k^* de clusters.

Plusieurs indices de validité de clusters sont proposés dans la littérature. Bezdek a défini deux indices: le Coefficient de partition (V_{PC}) [3] et l'Entropie de Partition (V_{PE}) [4]. Ils sont sensibles au bruit et à la variation de l'exposant m. D'autres indices V_{FS} et V_{XB} sont proposés respectivement par Fukayama et Sugeno [7] et Xie-Beni [18]; V_{FS} est sensible aux valeurs élevées et basses de m, V_{XB} donne de bonnes réponses sur un large choix pour c=2,…,10 et $1<m\leq 7$. Cependant, il décroît rapidement avec l'augmentation du nombre de clusters. Kwon et al. [16] ont apporté une amélioration à cet indice. Maria Halkidi et al. [10] ont défini V_{S_Dbw} basé sur les propriétés de compacité et de séparation de l'ensemble des données. Cet indice donne de bons résultats en cas de classes compactes et bien séparées, notamment quand il n'y a pas de chevauchement. Do-Jong Kim [14] a proposé un nouvel indice V_{SV}, en se basant sur la sommation des deux fonctions sous-partitionnement et sur-partitionnement. D'après le même auteur [14], cet indice s'est avéré plus performant que tous les autres cités dans ce paragraphe. Pour cela, dans notre partie expérimentale (section 4), nous nous limiterons à comparer notre nouvel indice V_{MEP} à V_{SV}.

Plus récemment, un nouvel indice de validité V_{OS} proposé par Dae-Won Kim et al en 2004 [15], exploite une mesure de séparation et une mesure de chevauchement entre clusters. Il est défini comme le rapport entre le degré de chevauchement et de séparation. La mesure du degré de chevauchement entre les clusters est obtenue en calculant le degré de chevauchement inter clusters. La mesure de séparation est obtenue en calculant la distance entre les clusters. D'après les auteurs [15], l'indice V_{OS} est plus performant que plusieurs autres indices. Cependant, il reste incapable de déterminer le nombre réel de clusters dans l'exemple des Iris [15], où il y a un réel chevauchement.

3. Nouvel indice de validité proposé V_{MEP}

3. 1. Principe du maximum d'entropie

Considérons un ensemble de données avec k clusters $c_1 \ldots c_j \ldots c_k$, et leurs centres respectifs $g_1 \ldots g_j \ldots g_k$. On définit les probabilités P_{ij} comme le lien entre le point i de sa classe c_j (j obtenu préalablement par l'algorithme de FCM) et son centre g_j. Les points i qui n'appartiennent pas à la classe c_j, ne possèdent aucun lien avec g_j ; c'est-à-dire $P_{ij}=0$.

On a : $$\sum_{i \in c_j} P_{ij} = 1 \; pour j = 1 .....k \qquad (1)$$

Pour toutes les classes, on obtient :

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

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

On définit une entropie qui mesure l'information apportée par toutes les classes par :

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

$$S = -\frac{1}{k} \sum_{j=1}^{k} \sum_{i \in c_j} P_{ij} \ln(P_{ij}) + \ln(k) \qquad (5)$$

$$S = \frac{1}{k} \sum_{j=1}^{k} S_j + \ln(k) \qquad (6)$$

Avec : $$S_j = -\sum_{i \in c_j} P_{ij} \ln(P_{ij}) \qquad (7)$$

S_j est l'entropie correspondant à la classe j. Le nombre optimal de classes k^* sera celui pour lequel l'entropie S est maximale.

3. 2. Calcul des coefficients P_{ij}

Pour chaque classe c_j, nous favorisons les points i les plus proches de son centre g_j en introduisant une contrainte additionnelle qu'on cherchera à minimiser, définie par :

$$W = \sum_{j=1}^{k} \sum_{i \in c_j} P_{ij} \left\| x_i - g_j \right\|^2 \qquad (8)$$

où $\| \ \|^2$ est la distance euclidienne.

Nous cherchons ainsi à avoir une concentration la plus élevée possible autour du centre g_j de chaque classe c_j. Maximiser S et minimiser W revient à minimiser l'expression suivante :

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

$$T = \frac{1}{k}\sum_{j=1}^{k}\sum_{i\in c_j} P_{ij}\ln(P_{ij}) + \ln(k) + \sum_{j=1}^{k}\sum_{i\in c_j} P_{ij}\left\|x_i - g_j\right\|^2 \quad (10) \quad \text{sous contrainte } \sum_{i\in c_j} P_{ij} = 1 \text{ ; pour j=1..k}$$

Le lagrangien de l'optimisation de la formule (10) sous les k contraintes est donné par :

$$L = \frac{1}{k}\sum_{j=1}^{k}\sum_{i\in c_j} P_{ij}\ln(P_{ij}) + \ln(k) + \sum_{j=1}^{k}\sum_{i\in c_j} P_{ij}\left\|x_i - g_j\right\|^2 + \sum_{j=1}^{k}\alpha_j\left(\sum_{i\in c_j} P_{ij} - 1\right) \qquad (11)$$

Où α_j est le multiplicateur de Lagrange associé à la j^{ème} contrainte. L'annulation de la dérivée de L par rapport à P_{ij} donne :

$$\frac{1}{k}\ln(P_{ij}) + \frac{1}{k}\left\|x_i - g_j\right\|^2 + \alpha_j = 0 \qquad (12)$$

Les expressions des P_{ij}, pour $i \in c_j$ et j = 1...k, sont déduites à partir de l'équation (12) par :

$$P_{ij} = Z_j^{-1}\exp\left[-k\left\|x_i - g_j\right\|^2\right] \qquad (13)$$

Tenant compte de la contrainte (1), Z_j est le coefficient de normalisation. En remplaçant P_{ij} par sa valeur dans (13), nous obtenons :

$$Z_j = \sum_{i\in c_j} \exp\left[-k\left\|x_i - g_j\right\|^2\right] \qquad (14)$$

Et par suite, les coefficients P_{ij} sont donnés par :

$$P_{ij} = \frac{\exp\left[-k\left\|x_i - g_j\right\|^2\right]}{\sum_{i\in c_j}\exp\left[-k\left\|x_i - g_j\right\|^2\right]} \qquad (15)$$

Cette expression ressemble à celle donnée par Xie-Beni [20]:

$$P_{ij} = \frac{\exp\left[-\beta D(x_i, g_j)\right]}{\sum_{i=1}^{N}\exp\left[-\beta D(x_i, g_j)\right]} \qquad (16)$$

Où $$D(x_i, g_j) = \sum_{k=1}^{p}\left|x_{ik} - g_{jk}\right|$$

est la norme 1 de la différence entre les deux vecteurs xi et le centre gj de la classe j ; β est un paramètre inconnu. Les deux expressions pour les coefficients P_{ij} dans (15) et (16) sont très similaires. Le grand avantage dans (15) réside dans l'élimination du paramètre β, dont le choix pose un problème en classification non supervisée.

3. 3. Définition du nouvel indice de validité proposé : V_{MEP}

Finalement, notre indice V_{MEP} est défini comme une entropie par :

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

Où S_j est défini par (7) qui utilise les P_{ij} définis dans l'équation (15). Le nombre optimal k^* de clusters sera celui pour lequel la valeur de V_{MEP} est maximale.

4. Résultats expérimentaux

L'indice V_{SV} proposé par Do-Jong Kim et al [14] et comparé dans plusieurs publications aux indices V_{PC}, V_{PE}, V_{FS}, V_{XB}, V_K et V_{SV} a montré une grande performance par rapport à tous les autres cités. Cet indice a été aussi utilisé avec succès dans un travail antécédent de l'un des auteurs [17] pour trouver le nombre optimal de clusters utilisant le modèle de mélange des gaussiennes (Gaussian Mixture Mode : GMM), et l'algorithme EM pour le processus de groupement, permettant d'extraire la forme des régions dans les images de textiles couleurs. Par conséquent, nous comparerons notre nouvel indice V_{MEP} uniquement à V_{SV} sur des exemples de données artificiels et réels.

Pour tester la performance de notre nouvel indice V_{MEP}, nous l'utilisons pour déterminer le nombre optimal de clusters sur des données synthétiques et aussi sur les données réelles bien connues qui sont les Iris de Fisher. Partant des boules polonaises [5] générées selon des distributions normales dont les paramètres sont rapportés dans la Table-1, et présenté dans le premier graphique de la figure 1. On retrouve 4 clusters compacts, bien séparés et alignés sur une diagonale.

Nombre cluster	Nombre points	Moyennes	Covariances
Cluster 1	1000	(-4 ; -4)	$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$
Cluster 2	1000	(0 ; 0)	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
Cluster 3	1000	(4 ; 4)	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
Cluster 4	1000	(8 ; 8)	$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

Table 1 : *Paramètres utilisés pour générer BD1*.

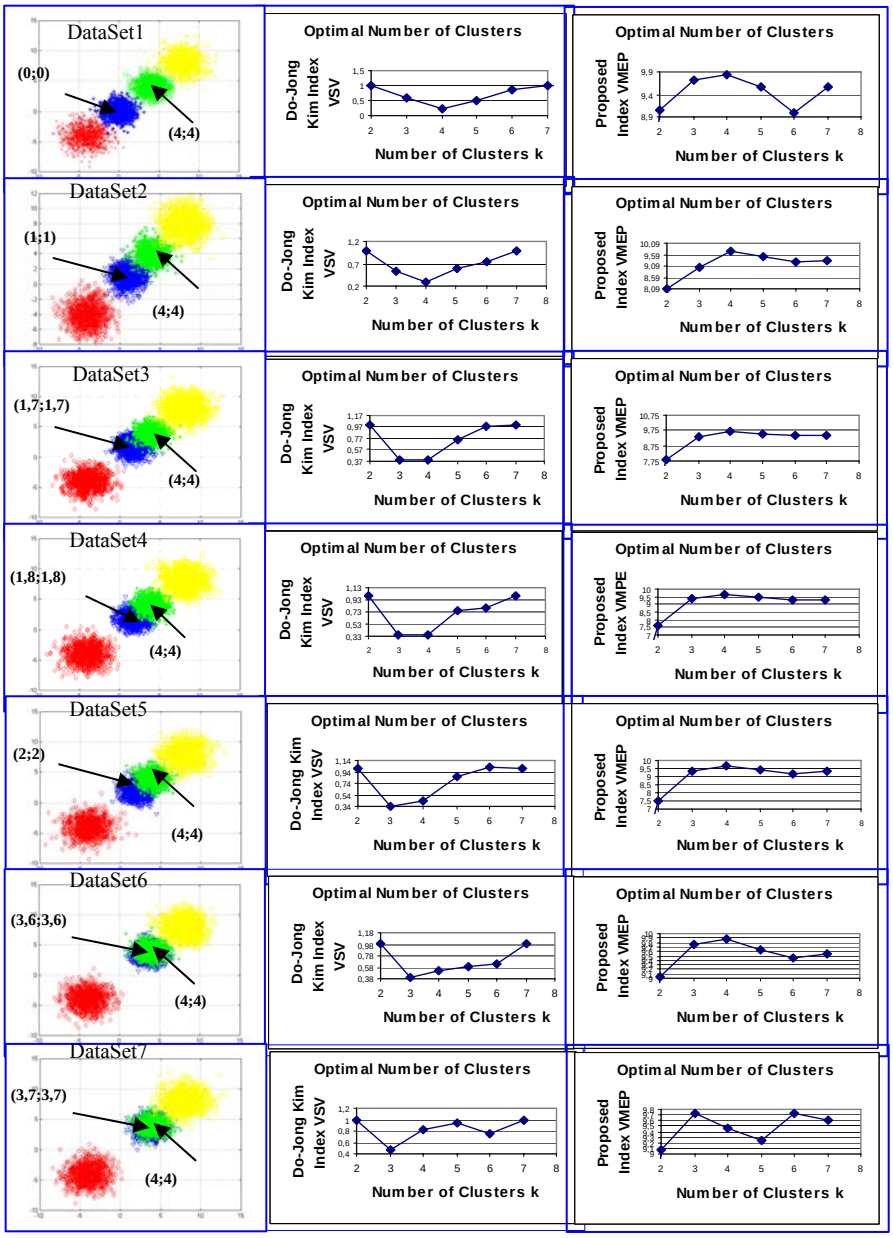

Figure1 : *Indice de Do-Jong Kim's V_{SV} (valeur minimale) et l'indice propose V_{MEP} (valeur maximale), affichés respectivement pour BD1, ... BD7.*

Les 15 autres bases de données nommés : DataSet2, DataSet3, DataSet4, DataSet5, DataSet6, DataSet7, DataSet8, DataSet10, DataSet11, DataSet12, DataSet13, DataSet14, DataSet15 et DataSet16, sont dérivés de la première en produisant un chevauchement croissant entre les deux clusters 2 et 3. On déplace les coordonnées du centre du cluster 2 initialisés à (0, 0) (table-1), en une série de coordonnées établies comme suit : (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), et finalement (4; 4) qui sont les coordonnées du centre du cluster 3 (table-1). Pour Dataset1, les deux clusters 2 et 3 sont complètement distincts et

pour Dataset16 ils sont pratiquement confondus. La Figure-1, la Figure-2, et la Figure-3, montrent les bases de données générées avec le chevauchement croissant des clusters 2 et 3.

Maintenant, nous appliquons V_{SV} et V_{MEP} à ces ensembles de données, notre indice de validité proposé V_{MEP} est-il plus efficace que V_{SV} **?** Si oui, jusqu'à quelle limite ?

Les résultats de validité de clustering obtenus en utilisant V_{SV} et V_{MEP} sont montrés dans la Figure-1. Pour DataSet1, où les deux clusters 2 et 3 sont bien séparés, les deux indices V_{SV} et V_{MEP} proposent 4 clusters ce qui est le nombre correct de clusters. Pour DataSet2, DataSet3 et, DataSet4, où débute un faible chevauchement entre les clusters 2 et 3, aussi V_{SV} et V_{MEP} sélectent correctement 4. Pour DataSet5, V_{SV} donne 3 clusters, ce qui est un nombre incorrect de clusters réellement existants. En augmentant le degré de chevauchement dans DataSet6, DataSet7, V_{SV} n'arrive plus à détecter le nombre correct de clusters, alors que V_{MEP} continu à donner 4 clusters pour tous ces ensembles de données (DataSet5, DataSet6, et DataSet7).

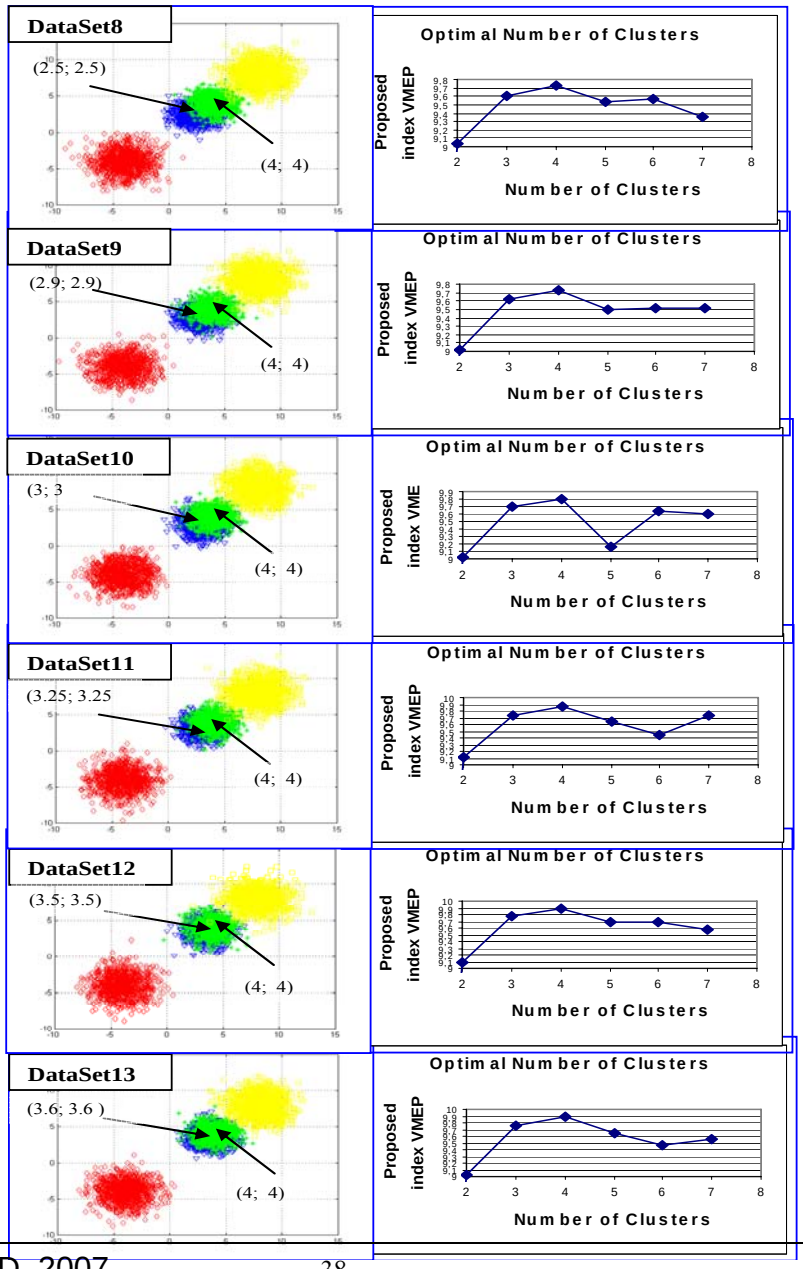

Figure 2. Results of clusters validation using the proposed V_{MEP}, displayed from DataSet8 to DataSet13

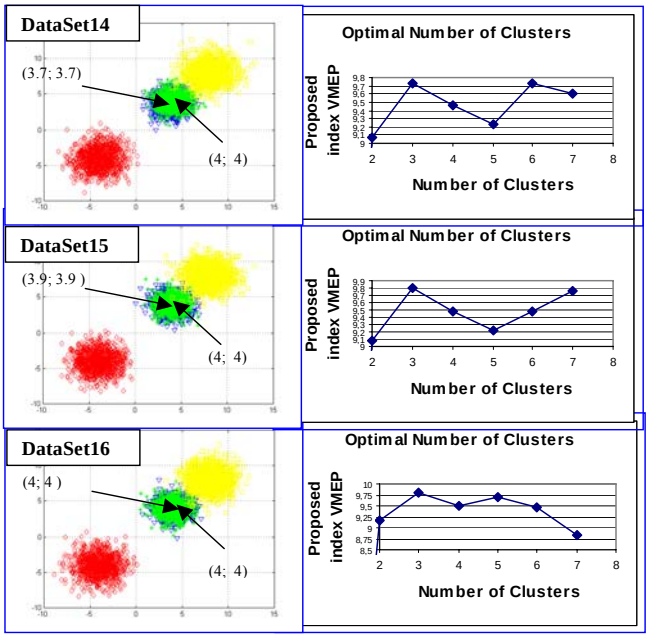

Figure 3. Results of clusters validation using the proposed V_{MEP}, displayed from DataSet14 to DataSet16

Vu ces résultats, on conclu que V_{SV} fonctionne bien seulement en présence d'un faible degré de chevauchement, et donne un nombre incorrect de clusters quand le chevauchement devient relativement plus important. Nous arrêtons alors d'appliquer V_{SV} pour les ensembles de données ayant un fort degré de chevauchement comme DataSet8…DataSet16 et nous continuons d'appliquer V_{MEP} pour voir jusqu'à quelle limite il peut être performant.

Les résultats obtenus par l'application de V_{MEP} à DataSet8…DataSet13, sont présentés respectivement sur la Figure-2, sur cette dernière, on peut voir que V_{MEP} peut encore bien fonctionner, il donne 4 comme nombre de clusters.

De DataSet14 à DataSet16, les coordonnées du centre du cluster 2 (les coordonnées du centre du cluster 3 restent fixées) sont respectivement : (3.7; 3.7), (3.9; 3.9), et (4; 4). Ces coordonnées deviennent très proches des coordonnées du centre du cluster 3 qui sont (4; 4). Le chevauchement devient alors très fort et dans ce cas, on peut voir dans la figure 3 que les deux clusters 2 et 3 sont semblables. V_{MEP} ne peut plus détecter 4 clusters comme nombre optimal, mais donne uniquement 3 clusters, ce qui peut être considéré comme un résultat normal puisque les deux clusters deviennent confondus.

La performance de notre indice V_{MEP} a été aussi examinée sur des données réelles [2].Il s'agit d'un ensemble de 150 d'iris constitué de 3 classes, nommées Setosa, Versicolor et Virginica, de 50 iris chacune. La majorité des indices récents proposés dans la littérature échouent à donner le nombre correct de clusters. Parmi les plus récents et performants indices, l'indice V_{OS} proposé par Dae-Won Kim et al [15] en 2004. Cet indice a la particularité de bien traiter les cas de chevauchement mais, comme le mentionnent ses

auteurs [15], V_{OS} n'arrive pas à détecter les 3 clusters dans le cas des IRIS mais 2 clusters, ce qui est un résultat faux.

Dans la figure 4, on présente les résultats obtenus en appliquant V_{SV} et V_{MEP}. Les deux indices sélectionnent correctement 3 comme nombre optimal de clusters. Ici V_{SV} fonctionne bien car le chevauchement entre les deux clusters est relativement faible.

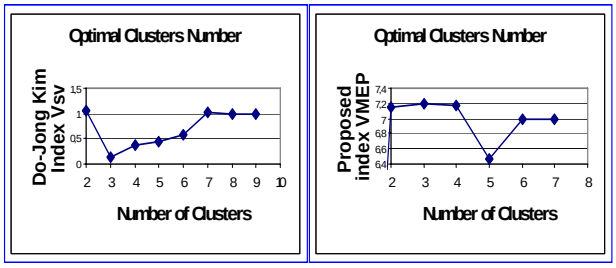

Figure4: Results of clusters validation using Do-Jong Kim's index V_{SV} (minimal value), and the proposed V_{MEP} (maximal value), applied to the Iris Data Set.

En conclusion la performance et la supériorité de notre indice proposé V_{MEP} est clairement établie comparativement à l'indice V_{SV}, et par conséquent à tous les autres cités auparavant. V_{MEP} sélectionne le nombre correct de clusters jusqu'à DataSet13 (Figure-2). De DataSet14 à DataSet16 (Figure-3), V_{MEP} ne peut plus détecter 4 clusters puisque le cluster 2 et 3 deviennent presque un seul cluster.

Jusqu'a maintenant et comme vérifié dans notre travail précèdent [1] V_{MEP} est performant pour les modèles gaussiens. Maintenant nous allons l'utiliser pour les modèles non gaussiens ce qui constitue une extension de notre dernier travail [1].

La Figure 5 montre les résultats obtenus en testant V_{MEP} sur des banana formes. Nous avons générés 4 banana formes nommées respectivement Banana set1, Banana set2, Banana set3 et Banana set4. Pour les 4 figures et pour les différentes formes, V_{MEP} détecte le nombre réel et correct de clusters.

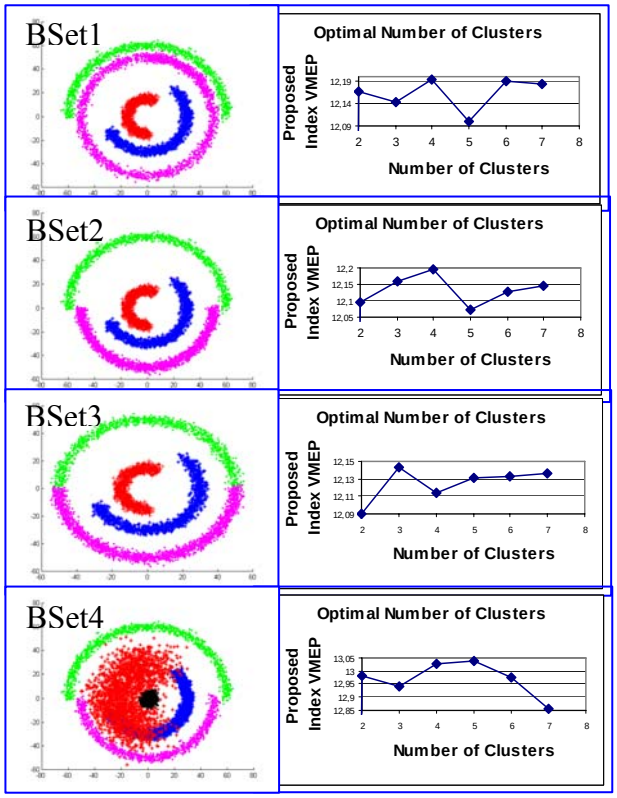

Figure 5*:* Results of clusters validation using V_{MEP} for some banana forms.

Banana set1 décrit 2 banana formes à l'intérieur d'un cercle enrobé lui même d'une banana forme. Pour cet ensemble Banana set1, V_{MEP} détecte 4 clusters, ce qui est le nombre correct de clusters pour cette figure. Pour banana set2, on garde les deux banana formes du centre mais cette fois à l'intérieur de deux symétriques banana formes avec un même centre mais ayant un rayon différent. Dans ce cas V_{MEP} retrouve 4 clusters, ce qui est encore le nombre correct de clusters pour banana set2. L'illustration de l'ensemble banana set3 montre deux banana formes à l'intérieur de deux symétriques banana formes avec un même centre et un même rayon. Pour les 3 premiers graphiques de la figure 5, on garde à l'intérieur les mêmes bananas formes. Pour le troisième graphique, V_{MEP} fonctionne encore bien et détecte 3 clusters, ce qui est le nombre correct.

Finalement, sur le dernier graphique de la figure 5, on teste notre indice sur une combinaison de différentes formes et chevauchement. Le résultat de cette application s'avère très intéressant, puisque V_{MEP} peut détecter 5 clusters, ce qui est le nombre correct. Ce dernier résultat complète la performance et la robustesse de notre nouvel indice.

5. Conclusion

Dans ce papier, nous avons proposé un nouvel indice pour l'évaluation de la qualité des résultats d'un algorithme de partitionnement. L'indice proposé, noté V_{MEP}, est basé sur le principe du maximum d'entropie. Le nombre optimal de clusters correspond au nombre k^* pour lequel l'indice V_{MEP} est maximal. La performance de notre nouvel indice est établie sur des exemples artificiels et réels. V_{MEP} peut détecter le nombre optimal correct de clusters

même avec un grand degré de chevauchement. Il peut être très utile dans les applications réelles en médecine, biologie, imagerie médicale,…où c'est important de connaître le nombre réel de clusters. Les résultats montrent la supériorité de notre indice V_{MEP} sur les autres. La performance de notre indice est aussi illustrée par sa capacité à bien fonctionner non seulement pour les modèles gaussiens, mais aussi sur d'autres formes non gaussiennes comme les banana formes, présentant des cas de chevauchement. Nous finirons par signaler un autre avantage de notre nouvel indice V_{MEP} : il n'est dépendant d'aucun paramètre produit par l'algorithme de classification utilisé ; de ce fait, il reste indépendant de l'algorithme de classification utilisé. Ceci nous donne la liberté de choisir celui qui semble le plus adapté pour l'application considérée; comme l'algorithme Gustafson-Kessel (GK) adapté pour les clusters de formes ellipsoïdales, ou encore l'algorithme EM. Ce sera l'objet d'un futur travail.

Références

[1] Ammor, O., Lachkar, A., Slaoui K., and Rais, N. "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.

[2] Anderson E., "The IRISes of the Gaspe Peninsula". Bull Am IRIS Soc, 59: 2–5, 1935.

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

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

[5] 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.

[6] Davies D. L. and Bouldin D. W. "Cluster Separation Measure", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 1, No. 2, pp. 95-104, 1979

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

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

[9] Halkidi, M., Batistakis, Y., Vazirgiannis, M., "Cluster Validity Methods: Part I", 2001.

[10] Halkidi, M., Batistakis, Y., Vazirgiannis, M., "Clustering Validity Checking Methods: Part II, 2002.

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

[12] Höppner, F., Klawonn, F., Kruse, R. and unkler, T., "Fuzzy Cluster Analysis-Methods for Classification: Data Analysis and Image Recognition". John Wiley & Sons, LTD, 1999.

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

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

[15] Kim, D. W., 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.

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

[17] 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 colour image segmentation method". Journal of the Textile Institute. Ref.: Ms. No. 10.1533.joti.2005.

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