Détermination du nombre de classes par le principe du maximum d'entropie pour des classes en chevauchement

A. LACHKAR[1], O. AMMOR[2], N. RAIS[3]
[1]E.S.T. M, Université Moulay Ismail, Meknès Maroc. E-mail: abdelmonaime_lachkar@yahoo.fr
[2]Laboratoire LMCS FSTF USMBA , Fès, Maroc, E-mail : w_ammor@yahoo.fr.
[3]Laboratoire ISQ. FSDM, USMBA, Fès, Maroc. E-mail : raissn@gmail.com

Résumé

Nous présentons un nouvel indice pour la détermination du nombre de classes basé sur le Principe du Maximum d'Entropie (V_{MEP}). La procédure est complètement automatique. Les performances de V_{MEP} sont illustrées à travers des exemples simulés et réels. Cet indice montre une grande robustesse, et une supériorité par rapport à d'autres méthodes existantes et récentes, particulièrement dans le cas du chevauchement spatial.

Mots clefs

Classification non supervisée, Principe du Maximum d'Entropie, chevauchement, nombre de 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 psychologie, la biologie, 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 chevauchement.

Plusieurs approches ont été proposées sur ce sujet pour différentes applications [1]-[7] 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]. Comme il est mentionné par Maria et al [10], les techniques basées sur le Critère Relatif citées dans la littérature [11]-[16], 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.

Dans cet article, nous présentons une nouvelle méthode de détermination du nombre optimal de classes d'un mélange multidimensionnel basée sur le principe du maximum d'entropie.

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 noté 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, particulièrement dans le cas du chevauchement spatial entre classes. On finira par la conclusion dans la section 5.

2 Indices de validité basés sur les critères relatifs

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}) [17] et l' Entropie de Partition (V_{PE}) [18]. 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 [19] et Xie-Beni [20]; 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. [21] ont apporté une amélioration à cet indice. Maria Halkidi et al. [15] 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 [22] a proposé un nouvel indice V_{SV}, en se basant sur la sommation des deux fonctions sous-partitionnement et sur-partitionnement. Cet indice s'est avéré plus performant que les autres cités auparavant.
Plus récemment, un nouvel indice de validité V_{OS} proposé par Dae-Won Kim et al en 2004 [23], 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 [23], 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 [23], 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 : $\sum_{i\in c_j} P_{ij} = 1 \; pour j = 1.....k$ (1)

Pour toutes les classes, on obtient :

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

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

$$S = -\sum_{j=1}^{k}\sum_{i\in c_j}\left(\frac{P_{ij}}{k}\right)\ln\left(\frac{P_{ij}}{k}\right) \quad (2)$$

$$S = -\frac{1}{k}\sum_{j=1}^{k}\sum_{i\in c_j} P_{ij}\,\ln(P_{ij}) + \ln(k) \quad (3)$$

$$S = \frac{1}{k}\sum_{j=1}^{k} S_j + \ln(k) \quad (4)$$

Avec : $S_j = -\sum_{i\in c_j} P_{ij}\,\ln(P_{ij})$ (5)

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 :

$$W = \sum_{j=1}^{k}\sum_{i\in c_j} P_{ij}\left\|x_i - g_j\right\|^2 \quad (6)$$

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 :

$$T = W - S \quad (7)$$

$$T = \frac{1}{k}\sum_{j=1}^{k}\sum_{i\in c_j} P_{ij}\,\ln\left(P_{ij}\right) - \ln(k) + \sum_{j=1}^{k}\sum_{i\in c_j} P_{ij}\left\|x_i - g_j\right\|^2 \quad (8)$$

sous contrainte $\sum_{i\in c_j} P_{ij} = 1$; pour j=1..k

Le lagrangien de l'optimisation de la formule (8) sous les k contraintes est donné par :

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

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 :

$$\frac{1}{k}\ln(P_{ij}) + \frac{1}{k} + \left\|x_i - g_j\right\|^2 + \alpha_j = 0 \qquad (10)$$

Les expressions des P_{ij}, pour i ∈ c_j et j = 1...k, sont déduites à partir de l'équation (10) par :

$$P_{ij} = \exp-\left(1 + k\alpha_j\right)\exp\left[-k\left\|x_i - g_j\right\|^2\right] \qquad (11)$$

Notons $Z_j = \exp\left(1 + k\alpha_j\right)$. Nous obtenons donc :

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

Tenant compte de la contrainte (1), Z_j est le coefficient de normalisation. En remplaçant P_{ij} par sa valeur dans (12), nous obtenons :

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

Et par suite, à partir de (12), les coefficients P_{ij} sont donnés par :

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

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 Pij définis dans l'équation (14). 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 [22] a été 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 [24] 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. Partant des boules polonaises [25] générées selon des distributions normales dont les paramètres sont rapportés dans la Table-1, nous avons générés 16 bases de données (BDi ; i=1..16) avec des degrés de chevauchement croissants entre les deux clusters 2 et 3 : pour la base de données BD1, les deux clusters 2 et 3 sont complètement distincts ; et pour BD16, ils sont pratiquement confondus. Etant donné le manque de place dans cet article, nous rapportons uniquement les figures au passage décisif. Ainsi, la figure-1 présente les 7 graphiques correspondant à BD1, BD2, BD5, BD6, BD7, BD13 et BD14. Dans BD1, on distingue 4 clusters compacts et bien séparés alignés sur la 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*.

Les BD2, BD5, BD6, BD7, BD13 et BD14 sont obtenues à partir de BD1 en déplaçant le cluster 2 ayant pour centre (0,0) (Table-1), vers le cluster 3 de centre (4, 4). Les centres respectifs du cluster 2 dans ces bases de données sont (1, 1), (1.7, 1.7), (1.8, 1.8), (2.0 ; 2.0), (3.6, 3.6), (3.7, 3.7).

Nous appliquons alors les indices V_{SV} et V_{MEP} sur l'ensemble de ces données.

Ces deux indices donnent le même résultat optimal, à savoir 4 clusters, tant qu'il s'agit d'un chevauchement léger. C'est le cas pour les trois premiers graphiques de la figure-1, dont les coordonnées des centres respectifs du cluster 2 sont (0 ; 0), (1 ; 1) et (1.7, 1.7).

A partir de BD6, dont le centre du cluster 2 est (1.8; 1.8), qui est relativement plus proche du cluster 3, c'est-à-dire présentant un chevauchement légèrement plus élevé, V_{sv} n'arrive plus à détecter le nombre de clusters corrects 4. Ceci est confirmé pour toutes les autres bases de données avec un chevauchement encore plus fort, notamment BDi, i=7…16.

Tandis que pour V_{MEP}, la limite de la bonne détection du nombre correct de clusters continue jusqu'à un très fort degré de chevauchement BDi, i=7…13 et illustré par les deux graphiques décisifs correspondant à BD7 et BD13 de la figure 1.

Pour BD14, BD15 et BD16, le nombre optimal de clusters déterminé par V_{MEP} est 3 comme illustré par le graphique

correspondant à BD14 de la figure-1. En effet, les deux clusters 2 et 3 sont pratiquement confondus.

La performance de V_{MEP} est montrée aussi par l'application aux données réelles Iris [26]. L'ensemble des données formé de 150 points répartis en 3 clusters de 50 points, nommés respectivement : Setosa, Versicolor, et Verginica. La plupart des récents indices cités auparavant n'arrivent pas à détecter le nombre réel de clusters des IRIS. Plus récemment, en 2004, Dae-Won Kim et al [23] a proposé un autre indice V_{OS} qui utilise le concept du degré de chevauchement et séparation. Cependant, il reste incapable de détecter le nombre réel de clusters en cas d'un chevauchement important, et comme mentionné par les auteurs dans [23], dans le cas des Iris, le nombre optimal de clusters qu'il détecte est 2, ce qui est un résultat faux. Dans la figure 2, nous présentons les résultats trouvés en utilisant V_{SV} et V_{MEP} sur les Iris. Les deux indices déterminent le nombre optimal correct de clusters qui est 3. Ici, V_{SV} fonctionne bien car il y a un faible degré de chevauchement.

Les bases de données générées artificiellement avec des degrés de chevauchement croissants (BDi, i=1…16), ainsi que les données réelles des Iris, nous ont permis de mettre en évidence les limites de performances des deux indices V_{SV} et V_{MEP}. La supériorité de V_{MEP} à V_{SV}, et par conséquent à tous les autres indices cités auparavant, est ainsi bien établie.

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, et ne nécessite aucun paramètre. 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, etc. où c'est important de connaître le nombre réel de clusters.

Les résultats trouvés montrent la supériorité de notre indice V_{MEP} sur les autres.

Notons que, comme tous les autres indices, V_{MEP} dépend aussi des résultats obtenus par l'algorithme FCM. Si celui-ci converge vers un minimum local, l'évaluation des indices de validités est inutile.

Nous finirons par signaler un autre avantage de notre nouvel indice V_{MEP} : il ne dépendant d'aucun paramètre produit par l'algorithme de classification utilisé ; de ce fait, il reste indépendant de l'algorithme de classification. 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] 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] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice Hall, New Jersey, 1988.

[3] R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis. New York: Wiley, 1973.

[4] J. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum, 198.

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

[6] J. Tou and R. Gonzalez, Pattern Recognition Principles. Reading, MA: Addison-Wesley, 1974.

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

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

[9] Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis Cluster Validity Methods : Part I

[10] Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis Clustering Validity Checking Methods: Part II

[11] J. C. Dunn: Well Separated Clusters and Optimal Fuzzy Partitions, Journal of Cybernetica, Vol. 4, pp. 95-104, 1974

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

[13] Subhash Sharma: Applied multivariate techniques, John Wiley & Sons, Inc., 1996

[14] M. Halkidi, Y. Batistakis and M. Vazirgiannis: On Clustering Validation Techniques, Journal of Intelligent Information Systems, Vol. 17, No. 2-3, pp. 107-145, 2001

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

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

[17] Bezdek,J.C.,1974.Numerical taxonomy with fuzzy sets. J. Math. Biology 1,57 –71.

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

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

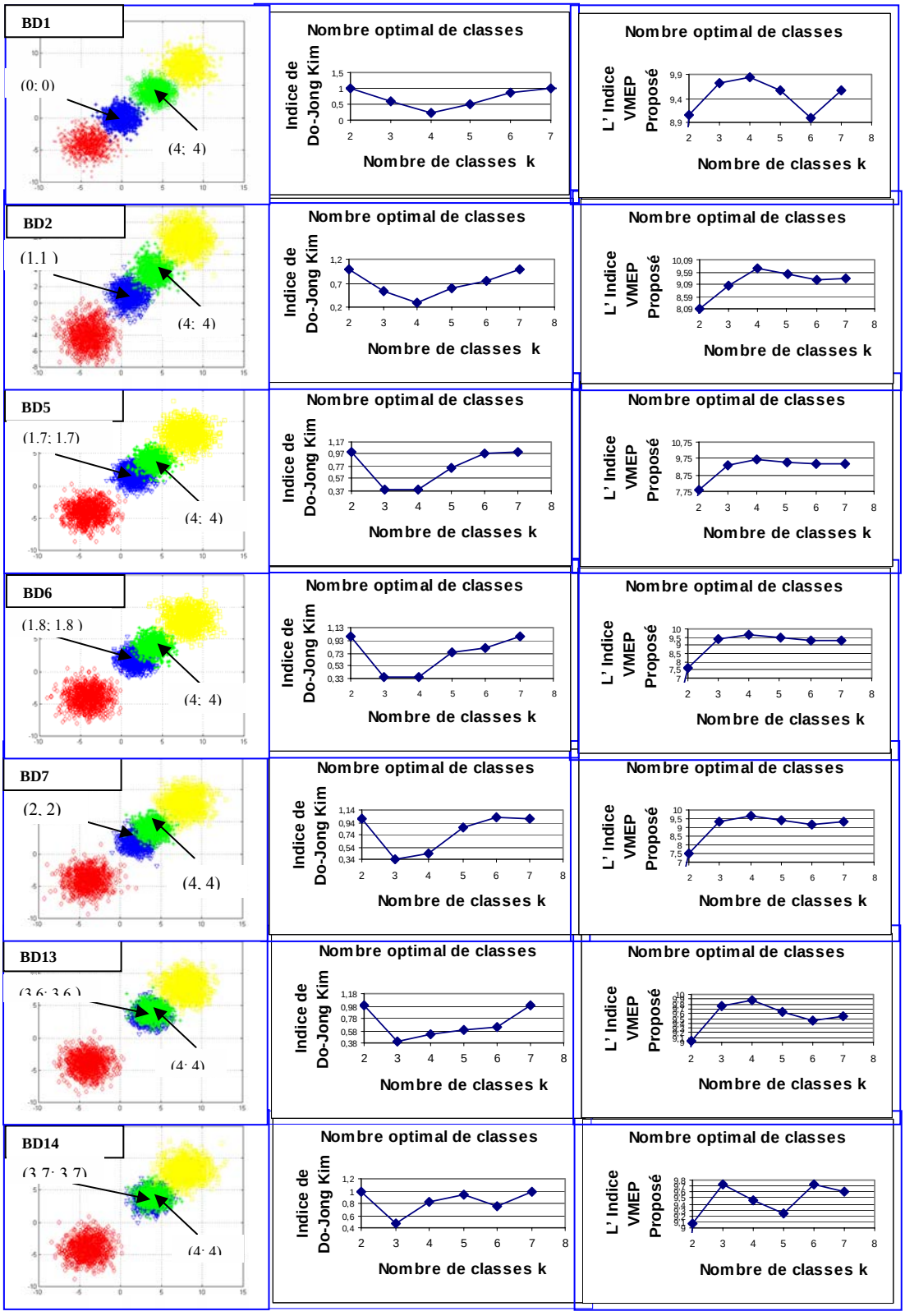

Figure1 : *Indice de Do-Jong Kim's V_{SV} (valeur minimale) et l'indice proposé V_{MEP} (valeur maximale), affichés respectivement pour BD1, BD2, BD5, BD6, BD7, BD13, et BD14.*

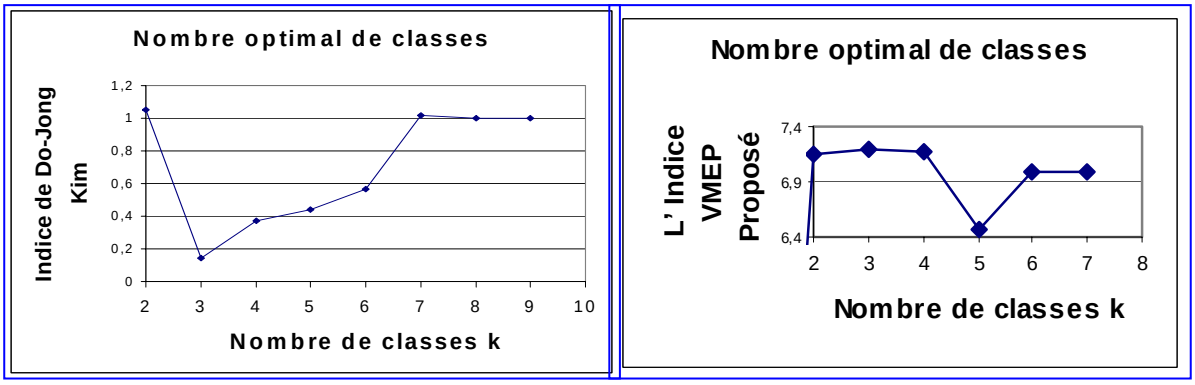

Figure2 *: Indice de Do-Jong Kim's V_{SV} (valeur minimale) et l'indice proposé V_{MEP} (valeur maximal) affichés respectivement pour les données Iris.*

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

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

[22] 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.

[23] Dae-Won Kim a ,Kwang H.Lee ,and Doheon Lee On cluster validity index for estimation of the optimal number of fuzzy clusters. Pattern Recognition. Vol 37. pp.2009 –2025. (2004)

[24] A. Lachkar, R. Benslimane, L. D'Orazio, E. Martuscelli,. A system for textile design patterns retrieval part 1: Design patterns extraction by adaptive and efficient colour image segmentation method. To appear in The Journal of the Textile Institute. Ref.: Ms. No. 10.1533.joti.2005.124R1

[25] 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.

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