ACP et ACI POUR LA REDUCTION DE DONNÉES EN IMAGERIE ASTRONOMIQUE MULTISPECTRALE

Farid FLITTI[1], Christophe COLLET[1,2], François BONNAREL[2]

[1]Université Strasbourg I, LSIIT : UMR CNRS 7005
LSIIT, Pole API, Bd S. Brant - 67400 Illkirch - France

[2]Observatoire astronomique de Strasbourg, UMR CNRS 7550
flitti@lsiit.u-strasbg.fr, collet@lsiit.u-strasbg.fr, bonnarel@astro.u-strasbg.fr

Résumé – Une technique de réduction de données astronomiques comme étape préliminaire à une classification d'images de galaxies lointaines est présentée. Le classifieur retenu est basé sur une modélisation markovienne causale en échelle sur le quadarbre. Ce papier présente les résultats encourageants obtenus par cette approche.

Abstract – A technique of astronomical data reduction as a preliminary stage of farway galaxies images classification is presented. The retained classifier is based on a in-scale-causal-Markovian modeling on the quadtree. This paper presents the encouraging results obtained by this approach.

1 Introduction

La segmentation non supervisée d'images reste à ce jour un problème difficile, en particulier dans le cas d'observations multispectrales ou hyperspectrales. Des travaux antérieurs ont permis de développer différents classifieurs dans le cadre bien établi de l'inférence bayésienne[1], en particulier de nombreux travaux basés sur une modélisation markovienne sur chaîne ou arbre[2, 3, 4] ont montré à la fois leur efficacité en temps de calcul et leur grande robustesse au bruit. Ces approches se distinguent d'autres techniques par la prise en compte de l'information de voisinage spatial et/ou en échelle, l'estimation non supervisée des paramètres du modèle et la capacité de s'adapter à des données multidimensionnelles [5]. Les cartes de segmentation obtenues sont satisfaisantes jusqu'à une dizaine de bandes, au delà les résultats se dégradent rapidement à cause du phénomène de Hughes[6] qui se manifeste par une perte de précision redoutable dans l'estimation des paramètres des mélanges de lois (terme d'attache aux données). Ainsi l'arrivée récente d'images hyperspectrales astronomiques à quelques centaines de bandes pose à nouveau le difficile problème de l'extraction de classes. Il s'agit alors de tirer profit de la grande quantité d'observations pour générer une carte de segmentation robuste tout en contournant le problème d'estimation de paramètres dans un espace de grande dimension à faible densité d'observations.

Ce problème, bien connu en télédétection, peut être abordé en effectuant au préalable une étape de réduction de l'espace de représentation, préliminaire à la classification[7]. En effet, la forte corrélation entre bandes spectrales adjacentes introduit une redondance d'information utilisée pour regrouper les images en sous ensembles de bandes fortement corrélées. Des projections linéaires locales sur chaque sous-ensemble [8, 9] peuvent alors être réalisées, et correspondre à une projection globale non linéaire prenant en compte les linéarités locales et minimisant l'espace de recherche des axes de projection.

Plus précisément, l'Analyse en Composantes Principales (ACP) est une technique classique utilisée en imagerie multispectrale astronomique (jusqu'à une dizaine de bandes) pour la recherche d'information utile [10]. Elle a pour objectif de décorréler les différentes composantes spectrales sur chaque pixel à l'aide d'une projection linéaire. C'est une méthode globale qui suppose implicitement que la distribution des données dans l'espace initial est un hyperellipsoïde caractérisé par la moyenne et la matrice de covariance globale [8]. Ainsi, cette approche basée sur un critère énergétique global, peut provoquer la perte définitive d'information caractérisant d'éventuels structures locales des données.

Plus récemment, l'Analyse en Composante Indépendantes (ACI) [11] a fait l'objets de nombreux développements. Cette technique considère les observations résultant d'une combinaison linéaire de sources indépendantes, supposées non gaussiennes. En effet, dans le cas gaussien la décorrélation entrainant l'indépendance, l'ACI n'a alors pas d'intérêt. Cette technique a été utilisée dans [12] pour l'analyse d'images astronomiques à 4 bandes.

Dans cet article, nous adoptons, en section 2, une stratégie de regroupement de bandes (cf. Fig. 1) en utilisant un algorithme de "bottom to up clustering" avec une mesure de similarité multiéchelles[13]. La réduction au sein de chaque sous ensemble sera alors réalisée (section 3) par une ACP ou une ACI (algorithme FastICA avec décorrélation déflationniste [11]). Les images ainsi réduites alimentent un classifieur markovien hiérarchique défini sur une structure de type quadarbre[4]. Cette stratégie a été validée sur images de synthèse (9 bandes) et testée sur images astronomiques (12 bandes). Les résultats obtenus sont présentés dans la quatrième partie de l'article.

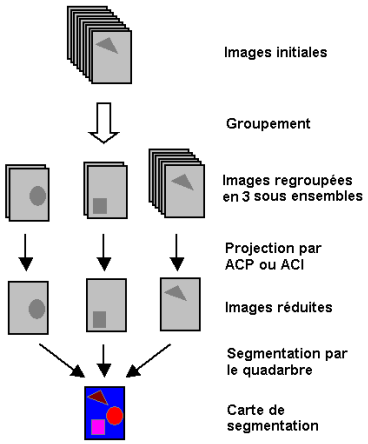

FIG. 1 – Technique de réduction proposée.

2 Regroupement des bandes spectrales

Le regroupement des différentes bandes spectrales est basé sur une technique de coalescence de bas en haut (”bottom to up clustering”) utilisant une métrique de similarité multiéchelle basée sur les histogrammes normalisés [13] auquel sont ajoutés les barycentres. Le barycentre g d'une image est défini par ses coordonnées x_g et y_g données par :

$$(x_g, y_g)^t = \frac{1}{\sum_x \sum_y f(x,y)} \sum_x \sum_y (x,y)^t f(x,y) \quad (1)$$

où x et y sont les coordonnées d'un pixel dans l'image et $f(x,y)$ la luminosité de ce pixel.

Chaque image est décomposée en échelle grâce à une technique de filtrage/décimation. La distance entre deux images s'évalue en sommant les différences entre leurs histogrammes et barycentres sur toutes les échelles. Nous obtenons ainsi une matrice $\{d_{ij}\}$ des distances entre les images i et j, $i, j \in \{1, \cdots, N\}$, où N désigne le nombre de bandes. Comme $d_{ij} = d_{ji}$, seule la partie triangulaire supérieur de $\{d_{ij}\}$ est stockée. Soient $C_i = \{i\}$, $i = 1, \cdots, N$, les N clusters initiaux et $S = \bigcup_i C_i$ l'ensemble de tout les clusters. La technique consiste à fusionner à chaque itération les deux clusters C_i et C_j dont la distance est minimale pour former un nouveau cluster C_k . Les clusters C_i et C_j sont éliminés et la distance entre C_k et les autres clusters est calculée. L'algorithme peut être décrit comme suit :

- $S \leftarrow \{C_1, \cdots, C_N\}$
- Pour chaque C_i, $C_j \in S^2$ calculer d_{ij}
- Pour chaque $K = N + 1$ à $2N - 1$
  {
  - $(C_{\hat{i}}, C_{\hat{j}}) = \arg\min_{(C_i, C_j) \in S^2} d_{ij}$
  - $C_k \leftarrow C_{\hat{i}} \bigcup C_{\hat{j}}$
  - $S \leftarrow \{S - \{C_{\hat{i}}\} - \{C_{\hat{j}}\}\} \bigcup C_k$
  - $\forall\, C_h \in S - \{C_k\}$ calculer : $d_{hk} \leftarrow d_{hi} + d_{hj} - d_{ij}$

  }

Il existe plusieurs façons de calculer d_{hk}, celle retenue dans l'algorithme précédent donne la structure en arbre la plus équilibrée pour les données[14].

3 Réduction et classification

L'ACP recherche une transformation W qui décorrèle les composantes multispectrales dans l'espace de projection selon l'équation :

$$\mathbf{z} = W\mathbf{y} \quad (2)$$

avec $\mathbf{y}$ le vecteur des données observées de dimension N, W la matrice de projection et $\mathbf{z}$ les vecteurs projetés. La matrice de covariance des données est calculée et décomposée en N valeurs propres par ordre décroissant, seules les m premières sont retenues. Les vecteurs colonnes de la matrice de projection sont donnés par les vecteur propres correspondants aux valeurs propres retenues. Nous obtenons ainsi une approximation optimale au sens de l'erreur quadratique moyenne des données initiales sur un espace réduit[15].

L'ACI repose sur le modèle suivant [11] :

$$\mathbf{y} = A\mathbf{s} \quad (3)$$

avec A la matrice de mélange et $\mathbf{s}$ les sources indépendantes. L'objectif est d'estimer la matrice de projection $W = A^{-1}$ pour retrouver les sources supposées indépendantes et non gaussiennes. En se basant sur un corollaire du théorème de la limite centrale, qui affirme que la somme de deux variables aléatoires est plus proche de la gaussienne que n'importe laquelle des deux variables, la majorité des techniques d'ACI cherchent W maximisant la non gaussiannité des sources. Cette dernière est mesurée par le Kurtosis (cumulant normalisé d'ordre 4) peu robuste, ou la néguentropie définie par :

$$J(\mathbf{y}) = H(\mathbf{y_g}) - H(\mathbf{y}) \quad (4)$$

où $H(\mathbf{y})$ est l'entropie de $\mathbf{y}$ et $H(\mathbf{y_g})$ l'entropie d'un vecteur gaussien de même covariance. Hyvärinen a proposé une approximation robuste de la néguentropie et un algorithme rapide pour sa maximisation (FastICA) [11]. Nous avons choisi la version FastICA avec décorrélation déflationiste qui recherche les vecteurs colonnes de W l'un après l'autre en favorisant les premiers. Elle se base sur la procédure d'orthogonalisation de Gram-Schmidt qui contraint le vecteur colonne w_i à appartenir à l'espace orthogonale aux $i - 1$ vecteurs déjà determinés. Comme l'indépendance implique la décorrélation, la maximisation de la néguentropie se fait sur l'espace réduit des données décorrélées par une ACP initiale où toutes les valeurs propres sont gardées.

Pour les deux techniques, nous n'avons gardé qu'un seul axe de projection pour chaque sous ensemble établi par regroupement. Les images ainsi obtenues alimentent un classifieur basé sur une modélisation markovienne causale en échelle. La décision est effectuée selon le critère MPM (Mode de la Marginale *a Posteriori*) :

$$\hat{x}_{opt} = \arg_x \min \sum_{x \in \Omega_x} L(x, \hat{x}) P(Y = \mathbf{y}, X = x) \quad (5)$$

où $\mathbf{y}$ désigne les observations multispectrales et x est le champ des étiquettes (carte de segmentation) qui prennent leurs valeurs dans Ω_x, $P(Y = \mathbf{y}, X = x)$ est la distribution jointe et $L(x, \hat{x})$ est la fonction de coût :

$$L(x, \hat{x}) = \sum_{m \in M} (1 - \delta(x_m, \hat{x}_m)) \quad (6)$$

avec M l'ensemble de tous les sites du champ et $\delta()$ est la fonction de Kronecker.

l'estimation des paramètres est réalisée grâce à l'algorithme ICE [4].

4 Résultats

Les résultats obtenus sur images de synthèse 9-bandes (Fig. 2) sont présentés sur la Fig. 3 pour 3 et 6 regroupements. Les performances obtenues sur ces images pour l'ACP et l'ACI en fonction du nombre de regroupements ("clusters") sont présentées Fig. 4. Dans les deux cas on remarque que la courbe présente un minimum d'erreur en fonction du nombre de groupements, puis les résultats se dégradent pour deux raisons : 1) quatre classes présentes sur les images de synthèse (carré, triangle, cercle et fond) nécessitent idéalement 4 clusters ; 2) le phénomène de Hughes devient perceptible lorsqu'on atteint la dizaine de bandes. La figure 5 présente une observation de galaxies lointaines (située dans la zone "Chandra Deep Field south") à faible rayonnement sur 12 bandes spectrales. Les deux cartes de segmentation obtenues sur 4 classes avec une ACP ou ACI préalable à la segmentation sur le quadarbre markovien (image en bas à gauche et à droite respectivement de la figure 5) montrent dans le cas de l'ACI une meilleure détection des zones étendues à faible rapport signal sur bruit.

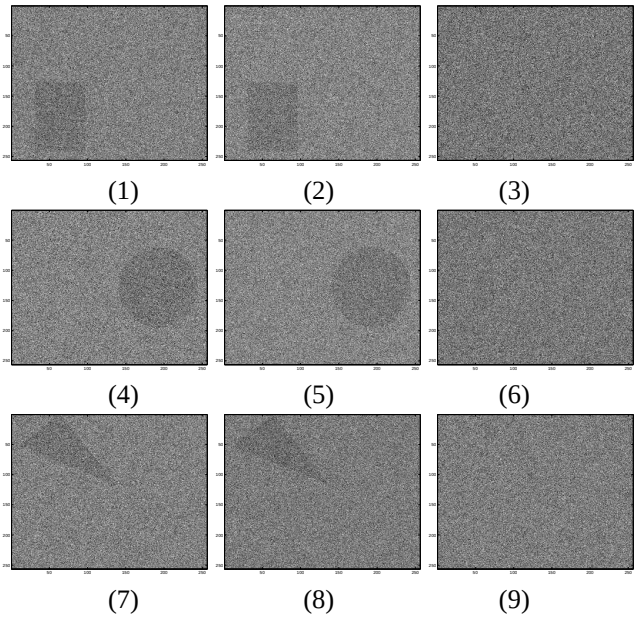

FIG. 2 – Les neufs images de synthèse utilisées.

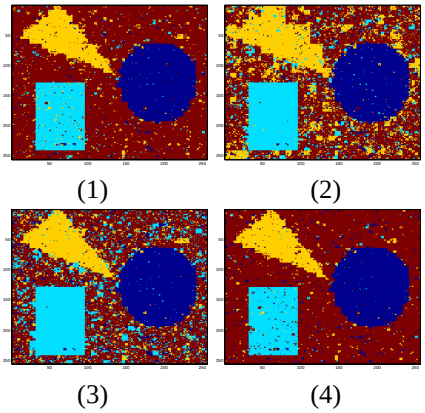

FIG. 3 – Les résultats des segmentations sur 4 classes pour les images de synthèse avec 3 clusters ($\{\{3,6,7,8,9\},\{1,2\},\{4,5\}\}$ Fig. 2) : (1) ACP (2) ACI , et 6 clusters ($\{\{3,9\}\{6\},\{7,8\},\{1\},\{2\},\{4,5\}\}$ Fig. 2) : (3) ACP (4) ACI .

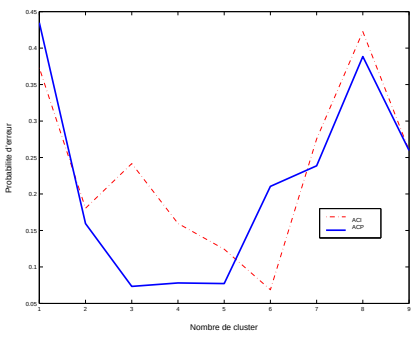

FIG. 4 – Comparaison de l'erreur de classification en fonction du nombre de clusters pour l'ACI et l'ACP.

Références

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley and Sons, 2001.

[2] J.M. Laferté, *Contribution à l'analyse d'images par modèles markoviens sur des graphes hiérarchiques. Application à la fusion de données multirésolution*, Thèse de doctorat, Université de Rennes 1, IRISA, Octobre 1996.

[3] N. Giordana, *Segmentation non supervisée d'images multispectrales par chaînes de Markov cachées*, Thèse de doctorat, Institut National des Télécommunications (INT), décembre 1996.

[4] J. N. Provost, *Classification bathymétrique en imagerie multispectrale SPOT*, Thèse de doctorat, Université de Bretagne Occidentale - Ecole Navale (Laboratoire GTS), Juin 2001.

[5] C. Collet, M. Louys, J. N. Provost, and A. Oberto, "Fusion of astronomical multiband images on a markovian quadtree," *Information Fusion, http ://www.fusion2002.org/*, July 2002, Annapolis, Maryland, USA.

[6] G.F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Information Theory*, vol. 14(1), pp. 55–63, 1968.

[7] David Landgrebe, *Information Processing for Remote sensing*, chapter Information Extraction Principles and Methods for Multispectral and Hyperspectral Image Data, World Scientific Publishing Co., Inc., 1999.

[8] L. Jimenez and D. Landgrebe, "High dimensional feature reduction via projection pursuit," *School of Electrical and Computer Engineering, Purdue University, West Lafayette in 47907-1285*, April 1995.

[9] G. Rellier, *Analyse de textures dans l'espace hyperspectral par des méthodes probabilistes*, Thèse de doctorat, Univ. Sophia Antipolis - INRIA, novembre 2002.

[10] J.-L. Starck, F. Murtagh, and A. Bijaoui, *Image Processing and Data Analysis : The Multiscale Approach*, Cambridge University Press, 1998.

[11] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley and Sons, 2001.

[12] A.Bijaoui and D.Nuzillard, "Blind source separation of multispectral astronomical images," *MPA/ESO/MPE Joint Astronomy Conf., Mining the Sky*, July 31 - August 4 2000, Garching, Germany.

[13] J. Y. Chen and P. Bouman, *Image Database Management Using Similarity Pyramids*, Ph.D. thesis, Perdue University, May 1999.

[14] G. N. Lance and W. T. Williams, "A general theory of classificatory sorting strategies 1. hierarchical systems," *The Computer Journal*, vol. 9, no. 4, pp. 373–380, February 1967.

[15] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.

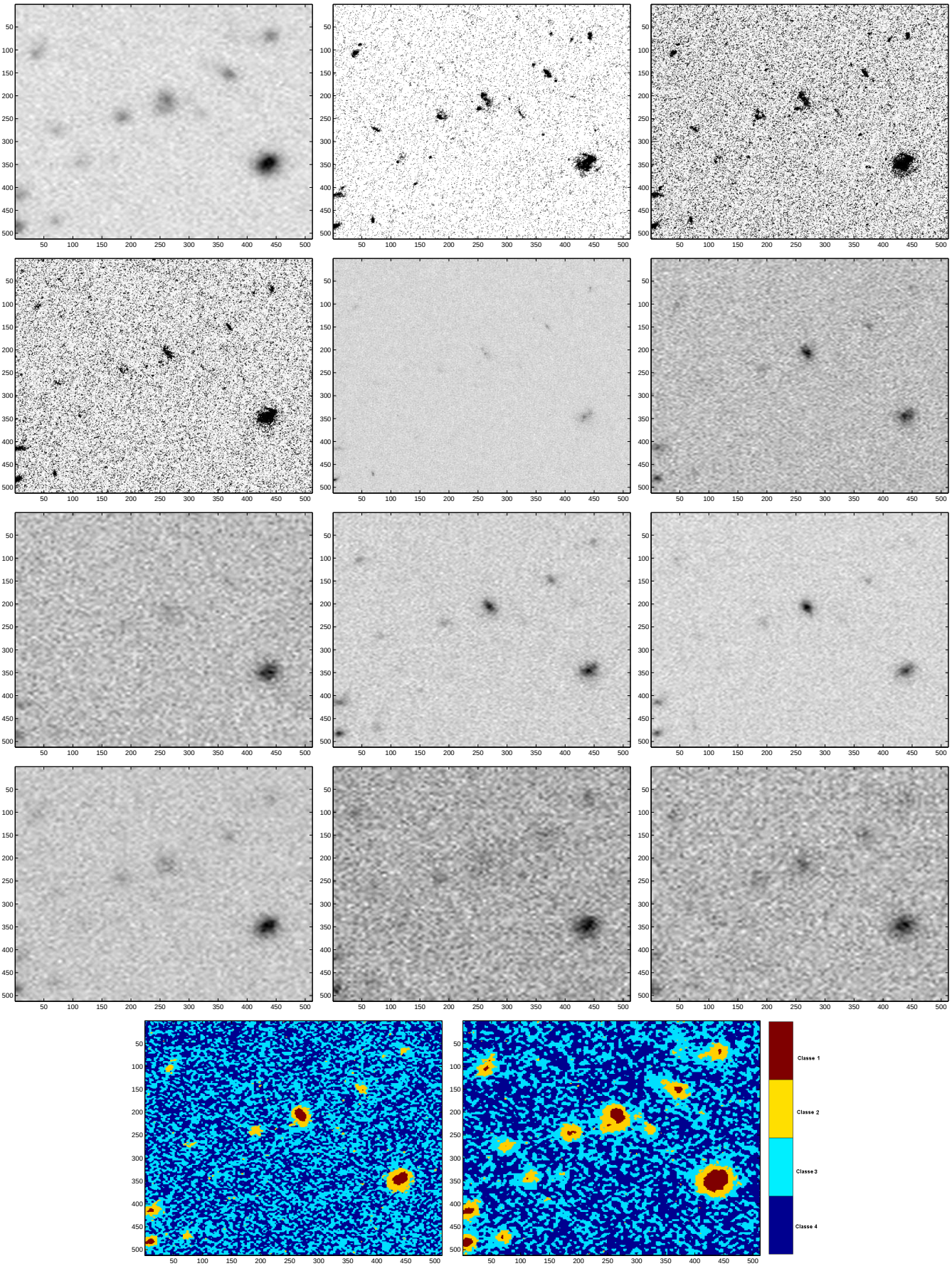

FIG. 5 – Les douze images astronomiques utilisées pour le test (en haut) avec les cartes de segmentations sur 4 classes avec 4 clusters en utilisant l'ACP et l'ACI (en bas de gauche à droite respectivement).