Utilisation de méthodes de classification hiérarchique pour une classification supervisée d'images satellitaires

Nadia OUARAB[1], Youcef SMARA[1], Jean-Paul RASSON[2]

[1]Laboratoire de Traitement d'Images, Institut d'Electronique
USTHB BP 32 El-Alia, Bab-Ezzouar 16111 Alger ALGERIE

[2] laboratoire GEOSATEL, FUNDP
8 Rempart de la vierge, Namur 5000 BELGIQUE

Résumé – La classification supervisée d'images satellitaires nécessite la connaissance et la nature des objets au sol. Le terrain est la seule référence fiable. Le problème se pose quand cette réalité n'est pas disponible. La solution que nous avons retenue est la sélection automatique de la base d'entraînement. La méthodologie adoptée consiste en la détermination des zones homogènes dans l'image par les tests statistiques de Wilcoxon et les supports des enveloppes convexes et le regroupement de ces zones par les méthodes de classification hiérarchique telles que celles du Single-linkage, Complete-linkage et Average-Linkage. Le classificateur du minimum de distance a été ensuite utilisé pour l'exploitation des différentes bases d'entraînement obtenues.

Abstract – The aim of this paper consists in detecting homogeneous regions in a satellite image. These regions could be used as a set of training data in the supervised classification. The supervised classification methods need the knowledge of earth landscapes and their nature. The ground is the only reference reliable for this task. The problem appears when this reality is not available. The solution we have adopted is the automatic selection of training data. We use nonparametric tests : univariate of Wilcoxon and an approach of supports comparison. For that purpose, we develop several hierarchical clustering methods, such as Single-linkage, Complete-linkage and Average-linkage.

1. Introduction

Pour une classification supervisée des images satellitaires, une étape préliminaire d'extraction d'échantillons d'entraînement des classes est nécessaire. Le principe [3] est de rechercher des zones homogènes dans l'image, de les regrouper par les méthodes de classification hiérarchique pour obtenir les classes d'apprentissage et de procéder par la suite à une classification supervisée d'images.

2. Génération des zones homogènes

Pour rechercher des zones homogènes dans l'image satellitaire, nous avons développé deux tests statistiques non paramétriques : le test univarié de Wilcoxon et le test multivarié des supports des enveloppes convexes [1][6].

2.1 Test univarié de WILCOXON

Nous considérons deux échantillons indépendants aléatoires $\tilde{V}_{n_x}=\{x_1,....,x_{n_x}\}$ et $\tilde{V}_{n_y}=\{y_1,....,y_{n_y}\}$ pris à partir des images tests d'entrée [1][5][6]. La combinaison de ces échantillons nous permet d'avoir N observations $(N=n_x+n_y)$ d'une même population ayant une densité inconnue. L'échantillon obtenu (ordonné et combiné) est désigné par un vecteur d'indicatrice $Z=(z_1,...,z_N)$ où :

$z_i=1$ si la ième variable aléatoire de l'échantillon ordonné est un élément de $\tilde{V}_{n_x}$.

$z_i=0$ si c'est un élément de $\tilde{V}_{n_y}$.

$\forall i=1,...,N$ avec $N=n_x+n_y$.

Le test de Wilcoxon est donné par :

$$W_x=\sum_{i=1}^{N} a_i z_i \qquad (1)$$

où les a_i sont des constantes appelées poids ou scores.

Un pixel de l'image appartient à une zone homogène si dans les quatre directions (horizontale, verticale, diagonale 45° et diagonale 135°) de l'image et pour tous les canaux, nous avons :

$$W_x \le W_\alpha \qquad (2)$$

où W_α est donné par les tables statistiques de Wilcoxon [5] pour chaque niveau de α.

2.2 Supports des enveloppes convexes

Soient $S_{\tilde{v}_{n_x}}$ et $S_{\tilde{v}_{n_y}}$ les enveloppes convexes des échantillons $\tilde{v}_{n_x}$ et $\tilde{v}_{n_y}$ [1][6].

Un indicateur du degré de similarité R entre les deux populations est défini par la relation suivante :

$$R = \frac{\left\{\tilde{V}_{n_x} \text{ I } S_{n_y}\right\} + \left\{\tilde{V}_{n_y} \text{ I } S_{n_x}\right\}}{n_x + n_y} \qquad (3)$$

avec :

- $\tilde{V}_{n_x}$ I S_{n_y} : nombre de pixels du second échantillon, présents dans l'enveloppe du premier échantillon.
- $\tilde{V}_{n_y}$ I S_{n_x} : nombre de pixels du premier échantillon, présents dans l'enveloppe du second échantillon.

Un pixel pris dans tous les canaux de l'image est homogène si dans les quatre directions nous avons :

$$R > R_\alpha \qquad (4)$$

où R_α est une valeur choisie par l'utilisateur.

Les deux tests statistiques définis précédemment, sont appliqués sur des images acquises par le satellite SPOT de la région de BLIDA (ALGERIE) de la figure 1. Celle-ci est de dimensions 256x256 pixels. Nous avons choisi de faire des tests pour des fenêtres 15x15. Le nombre b de pixels homogènes (nuage de points noirs) obtenu pour les figures 2 et 3 est de 1584 pour le test de Wilcoxon et de 1533 pour celui des supports des enveloppes convexes.

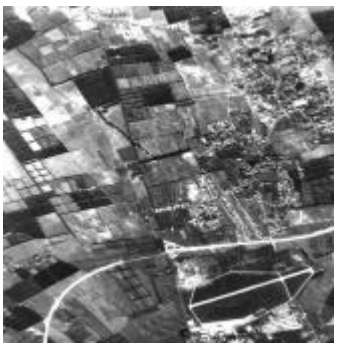


FIG. 1 : image de la région de Blida

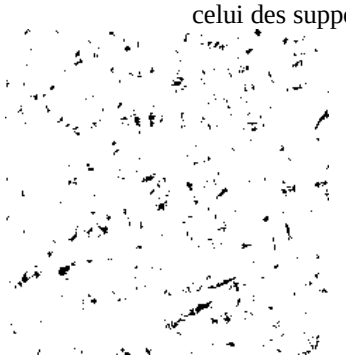


FIG. 2 : image de zones homogènes test de Wilcoxon (W_α=61)

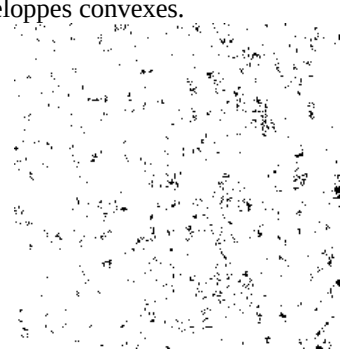


FIG. 3 : image de zones homogènes test des Supports (R_α=0.75)

3. Classification hiérarchique

Une fois les zones homogènes dans l'image déterminées, une méthode de classification automatique s'avère nécessaire. Celle-ci a pour but de partitionner l'image multispectrale en classes disjointes.

Il existe deux types de méthodes de classification hiérarchique [2][4], nous citons:

- La méthode hiérarchique descendante,
- La méthode hiérarchique ascendante.

L'étude effectuée est basée sur les méthodes hiérarchiques ascendantes. Celles-ci procèdent par des groupements successifs. Initialement, le nombre de pixels reflète le nombre de classes. Les pixels les plus similaires sont par la suite groupés jusqu'au nombre de classes choisies initialement. Le groupement des pixels se base sur le critère de la distance euclidienne.

Pour une classification hiérarchique ascendante [2][4] d'un ensemble de points, nous devons :

1/- Disposer d'un ensemble E de *b* éléments à classifier. b est le nombre de pixels homogènes à traiter,

2/- Chercher les deux éléments les plus proches que l'on agrège en un nouvel élément,

3/- Calculer les distances entre le nouvel élément et les éléments restants. On se trouve dans les mêmes conditions qu'à l'étape 1, avec seulement (b - 1) éléments à classer,

4/- Chercher de nouveau les deux éléments les plus proches, que l'on agrège. On calcule les nouvelles distances, et l'on réitère le processus jusqu'à obtention du nombre de classes.

4. Algorithmes de classification

Pour la mise en œuvre des méthodes de classification hiérarchique ascendante, nous disposons de plusieurs algorithmes dont ceux de KRUSKAL et PRIM [4].

4.1 Algorithme de KRUSKAL (1950)

Kruskal a proposé un algorithme [4] en 1950 qui nécessite :

- un ensemble *E* à classifier constitué de b pixels homogènes,
- une matrice de distance *D* obtenue à partir de l'ensemble *E*,
- un ensemble *A* vide au départ.

L'algorithme est défini par les étapes suivantes :

1/ - Ranger dans l'ordre croissant l'ensemble des distances euclidiennes $d(e_i,e_j)$, calculées précédemment. Une arête $e_i e_j$ est associée à chaque distance $d(e_i,e_j)$,

2/ - Mettre dans l'ensemble *A*, la première arête de la liste,

3/ - Mettre dans l'ensemble *A* l'arête suivante sauf si un cycle peut être formé avec les arêtes qui sont déjà dans *A* ,

4/ - Arrêter le processus si le graphe (*E, A)* est connexe, sinon recommencer la troisième étape [7].

4.2 Algorithme de PRIM (1957)

L'algorithme de Prim [4] est basé sur l'Algorithme de Kruskal, sauf que l'algorithme de Prim permet de faire un classement séquentiel des distances et détermine en même temps les arêtes qui ne font pas le cycle ; alors que l'algorithme de Kruskal utilise un classement général par ordre croissant des $b.(b-1)/2$ distances.

Prim [4] a proposé un algorithme qui utilise :
- un ensemble *E* que l'on veut classifier,
- une matrice de distance *D* obtenue de l'ensemble *E*,
- deux ensembles *A* et *T* qui sont vides au départ.

Les étapes suivantes permettent de décrire l'algorithme de PRIM, qui consiste à :
1/ - Prendre un élément quelconque de *E* et le mettre dans l'ensemble *T*,
2/ - Déterminer $e_i \in T$ et $e_j \notin T$ tel que :

$$d(e_i,e_j)=\min\{d(e,e') \text{ telque } e\in T \text{ et } e'\notin T\},$$

3/ - Mettre e_j dans l'ensemble *T* et l'arête $e_i e_j$ dans l'ensemble *A*. Réitérer la deuxième l'étape tant que *T* est diffèrent de E [7].

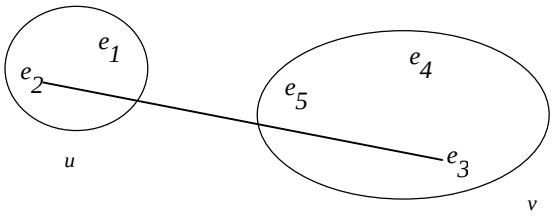
4.3 Distances de similitude

Les distances de similitude [4] utilisées dans la classification hiérarchiques sont :
- La méthode du lien unique (single-linkage),
- La méthode du lien complet (complète-linkage),
- La méthode du lien moyen (average-linkage),

et nous nous sommes intéressés aux deux dernières.

4.3.1 Méthode de lien complet

La méthode du lien complet consiste à :
1/ - Chercher la distance pour le plus proche couple de classes u et v. Soit d_{uv} cette distance,
2/ - Unir les classes u et v dans une classe (uv) et mettre à jour la matrice de distances,
3/ - Calculer les distances entre la classe (uv) et autres classes w par la relation : $d_{(uv)w}=\max\{d_{uw},d_{vw}\}$



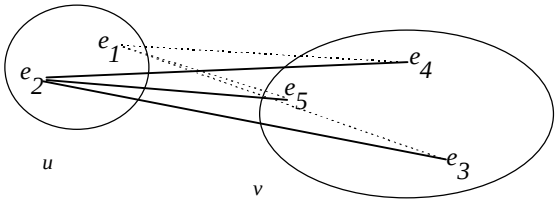
Méthode du lien complet (complete-linkage)
$d_{uv}=d_{e_2e_3}$

4.3.2 Méthode de lien moyen

Cette méthode [4] est similaire à la précédente, à l'exclusion de la troisième étape dont la distance de comparaison est basée sur la distance moyenne entre classes. Celle-ci est calculée de la manière suivante :

$$d(uv)w=\frac{\sum\sum d_{ij}}{\text{card}(uv)\,\text{card}(w)}$$

où $\sum\sum d_{ij}$ est une somme de distance d_{ij} entre les éléments de la classe w et la classe (uv),
card (uv) : nombre de pixels formant la classe uv,
card (w) : nombre de pixels formant la classe w.



Méthode du lien moyen (average-linkage)
$$d_{uv}=\frac{d_{e_1e_3}+d_{e_1e_4}+d_{e_1e_5}+d_{e_2e_3}+d_{e_2e_4}+d_{e_2e_5}}{6}$$

5. Implémentation algorithmique

Les différentes images obtenues par l'implémentation des méthodes du complete-Linkage et de l'average-linkage en utilisant l'algorithme de Prim sont données par les figures 5, 6, 7 et 8. La méthode que nous avons appliquée pour la classification d'images satellitaires en utilisant la base d'entraînement est la méthode du minimum de distance. Les images obtenues sont données par les figures 9, 10, 11 et 12.

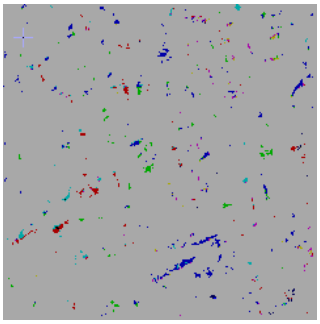


FIG. 6 : base d'entraînement (*complete-linkage, test de Wilcoxon*)

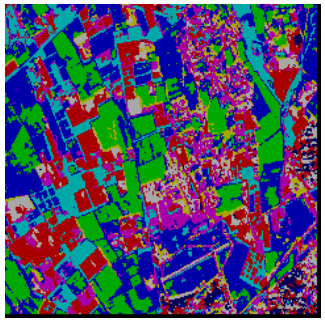


FIG. 9 : image classifiée (*complete-linkage, test de Wilcoxon*)

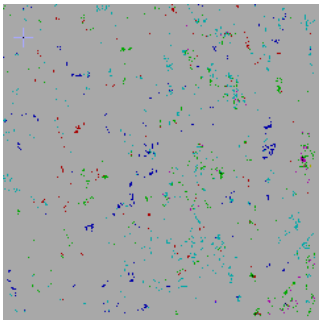


FIG . 7 : base d'entraînement *(average-linkage, test des Supports*)

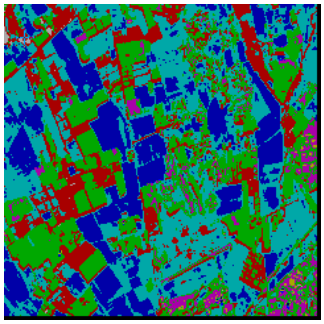


FIG . 10 : image classifiée *(average-linkage, test des Supports*)

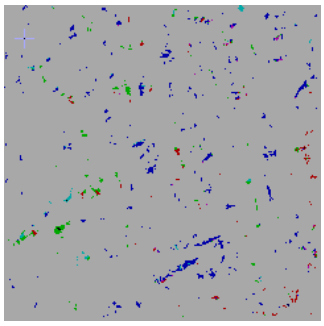


FIG. 8 : base d'entraînement *(average-linkage, test de Wilcoxon)*

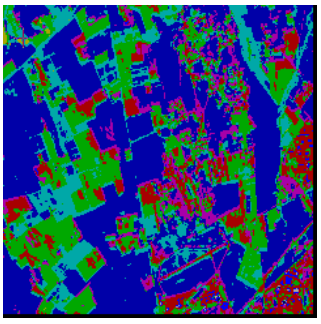


FIG. 11 : image classifiée *(average-linkage, test des Wilcoxon*)

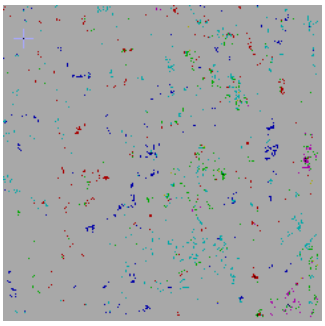


FIG. 5 : base d'entraînement *(complete-linkage, test des Supports*)

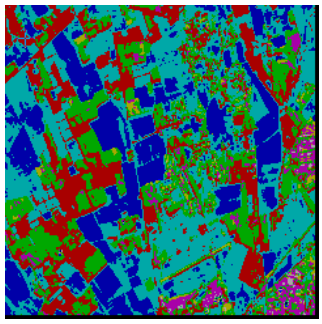


FIG. 12 : image classifiée *(complete-linkage, test des Supports*)

TAB. 1 : répartition des pixels homogènes en 10 classes

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	Total
figure 5	324	287	546	273	79	15	5	2	1	1	1533
figure 6	574	258	179	241	116	53	120	36	4	3	1584
figure 7	288	395	617	160	59	5	5	2	1	1	1533
figure 8	915	316	102	195	43	2	3	3	4	1	1584

TAB. 2 : répartition des pixels des images classifiées en 10 classes

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
figure 9	15903	10134	723	9117	7847	3369	3787	1098	62	16
figure 10	12063	11610	19838	11914	2027	847	186	15	21	43
figure 11	25862	12403	8450	5691	5489	189	39	13	417	11
figure 12	11883	14666	19515	9826	2153	186	236	15	41	43

6. Discussions et conclusion

Pour la détermination de zones homogènes dans l'image, les résultats obtenus dépendent du choix de la fenêtre de traitement ainsi que les paramètres relatifs à chaque test. Lors de l'implémentation des algorithmes de classification hiérarchique ascendante, un problème s'est posé concernant le temps d'exécution. Ce dernier dépend essentiellement du nombre *b* de pixels à classifier. Nous n'avons pas exposé les images obtenues par la méthode du single-linkage. Les résultats sont insuffisants, alors que ceux obtenus par les méthode du complete-linkage ainsi que l'average-linkage peuvent être améliorés en augmentant le nombre de pixels b à classer et en réduisant le temps de traitement en utilisant des algorithmes récents plus élaborés et rapides.

Références

[1] N. Ouarab. *Détermination automatique de la base d'entraînement et mise en œuvre de la méthode de classification d'images satellitaires.* (Thèse de magister); U.S.T.H.B, Alger, ALGERIE, 1997.

[2] A. Hardy. *Une nouvelle approche des problèmes de classification automatique : un modèle - un nouveau critère- des algorithmes- des applications.* Thèse de doctorat, facultés des sciences de Namur, Belgique, 1982.

[3] N. Ouarab, Y. Smara & J.-P. Rasson. *Détermination automatique de la base d'entraînement pour une classification supervisée d'images satellitaires par l'utilisation des tests statistiques.* Colloque international organisé par le cnes satellite-based : a tool for the study of the mediterranean, Tunis, 23-27 novembre 1998.

[4] L. Lebart, A. Morineau & J.-P. Fenelon. *Traitement des données statistiques, méthode & programmes.* Dunod, 1979.

[5] IMSL. *International Mathematical statistical libraries.* Inc. Houston, Texas(USA), 1979.

[6] V. Bertholet, G. Boudart & S. Lissoir. *Vers une automatisation des classifications sur bases de techniques de convexité.* Thèse de mathématique, Faculté des sciences de Namur, Belgique, 1995.

[7] E. Diday & J. Lemaire. *Eléments d'analyses de données. 1982.*