REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE ABOU-BEKR BELKAID - TLEMCEN

THÈSE

Présentée à :

FACULTE DES SCIENCES – DEPARTEMENT D'INFORMATIQUE

Pour l'obtention du diplôme de :

DOCTORAT EN SCIENCES

Spécialité: Informatique

Par :

Mme CHAOUCHE RAMDANE Lamia ép. STAMBOULI

Sur le thème

MULTI-CLASSIFIEURS DES IMAGES SATELLITAIRES

Soutenue publiquement le 18 Juin 2022 à Tlemcen devant le jury composé de :

CHIKH Azeddine	Professeur	Université de Tlemcen	Président
CHIKH Mohammed Amine	Professeur	Université de Tlemcen	Examinateur
FIZAZI Hadria	Professeur	Université d'Oran USTO	Examinatrice
CHOURAQUI Samira	Professeur	Université d'Oran USTO	Examinatrice
LAZOUNI Mohammed El Amine	MCA	Université de Tlemcen	Directeur
MAHI Habib	Maître de Recherche	CTS Arzew	Co-Directeur
EL HABIB DAHO Mostafa	MCA	Université de Tlemcen	Invité

Laboratoire de Recherche en Informatique de Tlemcen (LRIT)
BP 119, 13000 Tlemcen - Algérie

À la mémoire de mes chers parents,

À toute ma famille,

À mes enfants.

Remerciements

Une thèse est par essence un travail personnel. Cependant, sans un environnement propice et des soutiens constants, elle est un but inatteignable. Ainsi, j'aimerais remercier ceux sans qui ce travail n'aurait été possible.

Je voudrais tout d'abord adresser toute ma reconnaissance profonde à Monsieur Mahi Habib, Maitre de recherche et Directeur au Centre des Techniques Spatiales d'Arzew (CTS), pour avoir dirigé ce travail, pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui ont contribué à alimenter ma réflexion et pour les nombreux encouragements qu'il m'a prodiguée.

A Monsieur Mohammed El Amine Lazouni, maître de conférence à l'Université de Tlemcen , pour avoir accepté de dirigé cette thèse. Que ce travail soit le témoignage de mon profond respect et de ma reconnaissance.

A ceux de mes Professeurs qui me font l'honneur de siéger dans le jury, va ma gratitude :

Monsieur CHIKH Azeddine, Professeur à l'Université de Tlemcen. En hommage au grand honneur qu'il ma fait en acceptant de présider ce jury.

Monsieur CHIKH Mohammed Amine, Professeur à l'Université de Tlemcen. Qui a bien voulu me faire l'honneur de juger ce travail.

Madame FIZAZI Hadria, Professeur à l'Université des Sciences et de la Technologie d'Oran (U.S.T.O). Qui a bien voulu juger ce modeste travail.

Madame CHOURAQUI Samira, Professeur à l'Université des Sciences et de la Technologie d'Oran (U.S.T.O). Qui a bien voulu accepter d'être membre de jury.

Monsieur Mostafa El Habib Daho, Maitre de conférence à l'université de Tlemcen, pour son aide et ces précieux conseils avisés.

Je leurs exprime mes remerciements, les plus vifs.

Je remercie tous ceux qui ont contribué de prés ou de loin à l'aboutissement de ce travail.

Last but not least, mon mari, pour son soutien constant et ses encouragements. Merci.

Résumé

En télédétection, le clustering, également appelée classification non supervisée, est une tâche importante qui vise à partitionner une image donnée dans un espace multispectral en un certain nombre de classes spectrales (groupes), lorsque l'information in situ n'est pas disponible. Parmi les nombreux algorithmes de clustering existants, les plus utilisés sont le K-means, l'ISODATA, le FCM (Fuzzy C-Means), le SOM (Self Organizing Map) et plus récemment le K-Harmonic Means. Cependant, avec l'augmentation de la quantité de données détectées à distance et leur hétérogénéité, il devient difficile d'obtenir des résultats de clustering pertinents en utilisant un seul algorithme. De plus, chaque algorithme précité nécessite un certain nombre de paramètres et le plus important d'entre eux est le nombre de clusters, que l'utilisateur doit définir à priori.

Pour faire face à ces lacunes, les systèmes de classifications multiples (MCS), également connus sous le nom d'ensemble de clustering, est le consensus de différents algorithmes de clustering qui peut fournir la meilleure partition avec une grande précision et, par conséquent, surmonter les limites des approches traditionnelles basées sur des classificateurs uniques. Le MCS comprend deux étapes : la génération de partitions et la combinaison de partitions.

Dans cette thèse, nous étudions les avantages et les potentiels de cette technique dans le domaine de l'occupation du sol en utilisant différents types de données : Données synthétiques, données composites et données de télédétection. La première étape du MCS est assurée par quatre algorithmes de clustering, à savoir l'algorithme k-means, l'algorithme k-harmonic means (KHM), l'algorithme Bisecting K-means (BKM) et l'algorithme Self Organizing Map (SOM). Le meilleur clustering qui fait office de référence est obtenu selon l'indice WB. Les méthodes de ré-étiquetage et de vote sont utilisées dans la deuxième étape. Les résultats expérimentaux obtenus par le MCS surpassent légèrement les résultats du clustering individuel.

Mots clés : Clustering, K-means, k-harmonic means, Bisecting K-means, Self Organizing Map, indices de validité des clusters, données de télédétection.

Abstract

In remote sensing, clustering, also called unsupervised classification, is an important task that aims to partition a given image in a multispectral space into a number of spectral classes (clusters), when in situ information is not available. Among the many existing clustering algorithms, the most commonly used are K-means, ISODATA, FCM (Fuzzy C-Means), SOM (Self Organizing Map) and more recently K-Harmonic Means. However, with the increase in the amount of remotely sensed data and its heterogeneity, it becomes difficult to obtain relevant clustering results using a single clustering algorithm. Moreover, each algorithm mentioned above requires a number of parameters and the most important of them is the number of clusters, which the user has to define a priori.

To cope with these shortcomings, the Multiple Classifier System (MCS) is also known as ensemble clustering , is the consensus of different clustering algorithms can provide the best partition with high accuracy and consequently overcome limitations of traditional approaches based on single classifiers. The MCS involves two stages : the partitions generation and the partitions combination.

In this thesis, we investigate the potential advantages of this technique in the unsupervised land cover classification by using various kinds of data : Synthetic data, composite data and remotely sensed data. The first stage of the MCS is assumed by four clustering algorithms, the well-known k-means algorithm, the k-harmonic means algorithm (KHM), Bisecting K-means (BKM) and the self-organizing map (SOM). The best clustering is obtained according to WB index. The relabeling and the voting methods are used in the second stage. Experimental results obtained by the MCS outperform the results of the individual clustering.

Keywords : Clustering, K-means, k-harmonic means, Bisecting K-means, self-organizing map, cluster validity indices, remotely sensed data.

ملخص

في الاستشعار عن بعد ، يعتبر التجميع (clustering) ، الذي يُطلق عليه أيضًا التصنيف غير الخاضع للإشراف (classification non supervisée)، مهمة أساسية تهدف إلى تقسيم صورة معينة في مساحة متعددة الأطياف إلى عدد من الفئات (المجموعات) الطيفية ، عندما لا تتوفر المعلومات في الموقع.
من بين العديد من خوارزميات التجميع (clustering) الحالية الأكثر استخدامًا هي : K-mean، ISODATA، FCM (Fuzzy C-Means)، SOM (Self Organizing Map) ومؤخراً K-Harmonic Means. ومع ذلك ، مع الزيادة في كمية البيانات المكتشفة عن بعد وعدم تجانسها ، يصبح من الصعب الحصول على نتائج التجميع (clustering) ذات الصلة باستخدام خوارزمية واحدة.
بالإضافة إلى ذلك ، تتطلب كل خوارزمية سالفة الذكر عددًا من العوامل وأهمها عدد المجموعات التي يجب على المستخدم تحديدها مسبقًا .
للتعامل مع هذه النقائص ، فإن أنظمة التصنيف المتعددة (MCS) ، والمعروفة أيضًا باسم مجموعة التجميع (ensemble de clustering) ، هي توافق خوارزميات التجميع (clustering) المختلفة التي يمكن أن توفر أفضل تجزئة بدقة عالية وبالتالي التغلب على قيود الأساليب التقليدية القائمة على المصنفات الفردية. يتكون MCS من خطوتين: إنشاء التجزئات والجمع بين التجزئات .
في هذه الأطروحة ، ندرس مزايا وإمكانيات هذه التقنية في مجال الغطاء الأرضي باستخدام أنواع مختلفة من البيانات: البيانات التركيبية والبيانات المركبة وبيانات الاستشعار عن بعد. في المرحلة الأولى من MCS يتم الإعتماد على أربع خوارزميات تجميع (clustering)، وهي خوارزمية k-means ، خوارزمية (KHM) k-harmonic means، خوارزمية Bisecting K-mean (BKM) وخوارزمية خريطة التنظيم الذاتي Self Organizing Map (SOM).
إن أفضل تجميع مرجعي تم الحصول عليه وفقا لمؤشر WB. في الخطوة الثانية يتم استخدام طريقتي إعادة وضع العلامات (ré-étiquetage) والتصويت.
النتائج التجريبية التي تم الحصول عليها بواسطة MCS تجاوزت بشكل طفيف نتائج التجميع الفردي (clustering individuel) .

الكلمات المفتاحية : التجميع (clustering) ، K-means، k-harmonic means ، K-Bisecting means، خريطة التنظيم الذاتي (Self Organizing Map) ، مؤشرات صلاحية المجموعات (indices de validité des clusters)، بيانات الاستشعار عن بعد (données de télédétection).

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations et acronymes

ALOS	Advanced Land Observing Satellite
Alsat	Algerian satellite
ASAL	Agence Spatiale ALgérienne
BIC	Bayesian information Criterion
BKM	Bisecting k-Means
BMU	Best Matching Unit
BR	Basse Résolution
CSPA	Cluster-based Similarity Partioning Algorithm
CV	Cumulative Voting
DB	Davies-Bouldin
DECORATEs	Diversity Ensemble Creation by Oppositional Relabeling of Artificial Training Examples
EM	Expectation Maximization
EMR	ElectroMagnetic Radiation (rayonnement électromagnétique)
ENVISAT	ENVIronment SATellite
EQM	Erreur Quadratique Moyenne
ERTS	Earth Resources Technology Satellite
ESA	European Space Agency
FCM	Fuzzy C-Means
GMS	Geostationary Meteorological Satellite
GMM	Gaussian Mixture Models
GOES	Geostationary Operational Environmental Satellites
GPS	Global Positioning System
GTM	Generative Topographic Mapping
HBGF	Hybrid Bipartite Graph Formulation
HGPA	HyperGraph Partitioning Algorithm
HR	Haute Résolution
INSAT	Indian National Satellite System
IRS	Indian Remote Sensing
IS	Indice de Silhouette
ISODATA	Iterative Self-Organizing Data Analysis Technics A
ISS	International Space Station
KHM	K-Harmonic Means
K-PPV	K-Plus Proches Voisins
K- NN	K- Nearest Neighbours
MCS	Multiple Classifier System

MCLA	Meta-CLustering Algorithm
MLP	Multi Layer Perceptron (Perceptron multicouche)
MR	Moyenne Résolution
MS	Multi Spectral
MSS	Multi Specrtal Scanner
NASA	National Aeronautic and Space Administration
nm	Namomètre (un nanomètre = 10^{-9} mètre)
NMF	Nonnegative Matrix Factorization
PAN	PANchromatique
PV	Plurality Voting
PNN	Probabilistic Neural Network (le réseau de neurones probabiliste)
RBF	Radial Basis Function (Fonction de base radiales)
RF	Rotation Forest
Sentinel	Sentinelle en français
SOM	Self-Organizing Map
SPOT	Système Probatoire d'Observation de la Terre
SSW	Sum-of-Squares Within
SSB	Sum-of-Squares Between
TM	Thematic Mapper
THR	Très Haute Résolution
TIROS	Television Infrared Observation Satellite
UCS	Union of Concerned Scientists
USGS	United States Geological Survey
VAC	Voting Active Clusters
V-M	Voting-Merging
WPCK	Weighted Partition Consensus via Kernels

Introduction générale

Sommaire

La théorie dure tant qu'elle résiste à l'expérience ; elle se modifie et change le jour où elle est vaincue par les faits de l'expérience.

-CLAUDE BERNARD-

Contexte des travaux

Il existe de nombreuses méthodes et algorithmes disponibles pour le traitement des images satellites. Les plus connues sont le rehaussement, l'extraction de formes, la segmentation, la fusion, la détection de changements, la compression, la classification et la détection de formes.

Les images de télédétection sont largement utilisées en classification de la couverture terrestre, l'identification de cibles et la cartographie thématique, de l'échelle locale à l'échelle mondiale, en raison de leurs avantages techniques tels que la multi-résolution, la large couverture et les enregistrements multispectraux et hyperspectraux [Navalgund et al., 2007]. Avec les progrès de la technologie d'acquisition de données de télédétection, les images de télédétection peuvent être acquises par divers capteurs, par exemple, un spectromètre d'imagerie hyperspectral, des capteurs à haute résolution, un radar polarimétrique à ouverture synthétique, etc. [Zhang, 2010].

Une solution efficace consiste à générer un ensemble de classificateurs en combinant certains classificateurs individuels, que l'on appelle **système à classificateurs multiples**, **Multiple Classifier System** en anglais **(MCS)** ou ensemble de classificateurs [Steele, 2000][Briem et al., 2002][Smits, 2002][Benediktsson et al., 2007][Foody et al., 2007][Doan et Foody, 2007][Okun et Valentini, 2008][Waske et al., 2010]. Au cours des vingt dernières années, le MCS s'est développé rapidement et a été largement utilisé dans divers domaines tels que la reconnaissance des formes, le traitement des images et l'identification des cibles. En outre, le MCS, a été utilisé récemment par la société de télédétection, est considéré comme un moyen efficace d'améliorer les performances de classification des images de télédétection. De nombreux chercheurs ont étudié la possibilité de combiner plusieurs classificateurs pour obtenir un résultat satisfaisant [Briem et al., 2002][Lu et Weng, 2007].

Les systèmes à classificateurs multiples en télédétection

Les applications des MSC dans le domaine de la télédétection sont nombreuses et traitent principalement de la cartographie de l'occupation du sol et de la détection des changements. Ce qui suit un aperçu des différents travaux se trouvant dans la litérature.

[Giacinto et al., 2000] ont proposé une méthode basée sur la combinaison d'algorithmes neuronaux et statistiques (le classifieur naïf de Bayes, le perceptron multicouche, le réseau à fonctions de base radiales et le réseau de neurones probabiliste) pour la classification supervisée d'images de télédétection. Ils ont démontré que la combinaison d'algorithmes neuronaux et statistiques est un moyen efficace d'obtenir rapidement des valeurs de haute

précision après de courtes phases de conception et d'améliorer le compromis précision/rejet par rapport aux techniques individuelles.

[Wemmert, 2000] a proposé une méthode de clustering collaboratif appelée SAMARAH. Cette méthode consiste en la collaboration de différentes méthodes de clustering pour tenter de trouver un consensus sur le clustering d'un jeu de données. L'objectif de cette collaboration est de réduire l'impact du choix d'une méthode et de ces paramètres sur le résultat.

[Briem et al., 2002] ont étudié les performances de trois types de classificateurs multiples, à savoir les algorithmes de mise en sac (bagging algorithms), les algorithmes de boosting (boosting algorithms) et les classificateurs basés sur la théorie du consensus (consensus-theoretic classifiers), en termes de classification de données géographiques et de télédétection multisources. Les trois schémas ont donné de bons résultats et ont surpassé de plusieurs classificateurs uniques en termes de précision.

[Forestier et al., 2008] se sont intéressé à la collaboration des algorithmes de clustering dans le cadre de la classification orientée objet d'images de télédétection à très haute résolution des zones urbaines. La méthode permet d'effectuer une classification non supervisée multi-stratégique, donnant un résultat unique qui combine les résultats de toutes les méthodes de classification.

[Chi et al., 2009] ont appliqué un algorithme de classification d'ensemble, qui combine des modèles génératifs (mélange de gaussiens) et discriminatifs (machine à cluster de support) pour traiter le problème de quantité dans la classification des images hyperspectrales de télédétection. Les résultats ont permis d'améliorer la précision de la classification ainsi que sa robustesse.

[Han et al., 2012] ont proposé un algorithme hybride entre l'algorithme DECORATEs (Diversity Ensemble Creation by Oppositional Relabeling of Artificial Training Examples) et l'algorithme RF (Rotation Forest), afin d'éviter le sur-apprentissage dans la classification des images de télédétection. Les réseaux neuronaux à fonction de base radiale (RBF) ont été employés comme classificateurs de base. Dans les expérimentations, l'algorithme proposé a montré une précision de classification plus élevée et moins en sur-apprentissage.

Dans les travaux de [Su et Du, 2014], la stratégie diviser pour régner (divide-and-conquer strategy) a été proposée, dans laquelle un classificateur est appliqué à chaque groupe de bandes, et la sortie finale est le produit de nombreux classificateurs fusionnés ensemble.

Pour la segmentation des images hyperspectrales, [Berikov et al., 2014] ont présenté une nouvelle méthode de clustering d'ensemble. L'idée principale de la méthode est de réduire la quantité de données étudiées en utilisant une succession d'algorithmes k-means comme étape préparatoire. L'efficacité des algorithmes proposés est démontrée par des résultats de clustering sur des images hyperspectrales réelles.

[Sublime et al., 2016] ont proposé une approche originale basée sur le Generative Topographic Mapping (GTM) qui consiste à l'utiliser sur différents jeux de données où des clusters similaires peuvent être trouvés (mêmes

espaces de caractéristiques et distributions de données similaires). Les résultats des expériences révèlent que ce cadre est assez bon pour améliorer le regroupement final des cartes dans le processus collaboratif.

[Tatarnikov et al., 2017] ont proposé un algorithme de la moyenne des centroïdes. En utilisant un ensemble de partitions créées avec une approche basée sur les centroïdes, l'algorithme peut produire la partition consensuelle d'un ensemble de données en clusters.

[Miao et al., 2018] ont proposé un nouveau MCS pour résoudre le problème de la classification de l'occupation du sol avec des images de télédétection. Le système tient compte à la fois de l'efficacité et de l'efficience en combinant la taille et l'ensemble des classificateurs en même temps, ce qui est réalisé en transférant ces tâches à un problème d'optimisation. Les résultats expérimentaux montrent que le système MCS proposé peut classifier avec succès les images de télédétection avec une grande précision tout en réduisant le coût de calcul.

Malgré l'abondance des applications des MCS au traitement des données de télédétection, il convient de mentionner que la majorité des travaux cités ci-dessus traitent le cas de la classification supervisée et que seules quelques études ont été consacrées au cas non supervisé. Ceci est principalement dû au fait que l'étude de la relation précision-diversité des classificateurs de base dans le cas supervisé est difficile. Par conséquent, afin d'obtenir un résultat de classification unique et utile, cette recherche étudie l'utilisation d'une combinaison de méthodes de classification non supervisée.

Problématique

Ensemble de clustering est le processus qui consiste à combiner les multiples résultats de clustering d'un ensemble d'objets en un seul clustering amélioré. Il est parfois appelé solution consensuelle ou agrégation de clustering. Ces dernières années, diverses études ont été menées pour développer des méthodes d'ensemble de clustering inspirées par le succès de la méthode d'ensemble dans le domaine de l'apprentissage supervisé [Strehl et Ghosh, 2002][Fern et Brodley, 2004][*Topchy*[a] et al., 2004][Topchy et al., 2005]. Cependant, par rapport à la recherche sur les méthodes d'ensemble de classification, la construction d'un ensemble de clustering n'est pas simple, et des travaux supplémentaires sont nécessaires dans ce domaine.

Plusieurs raisons rendent la construction d'un ensemble de clustering plus difficile que celle de la classification supervisée. L'une d'entre elles est que le clustering est un apprentissage non supervisé dans lequel les données ne sont pas étiquetées. Il n'y a donc pas de connaissances préalables avec lesquelles l'algorithme peut découvrir la véritable structure de cluster, et il n'y a pas de "vérité terrain" pour valider le résultat du clustering. De plus, aucune technique de validation ne peut être utilisée pour régler les paramètres de l'algorithme de clustering. Il n'existe donc pas de lignes di-

rectrices permettant à l'utilisateur de sélectionner l'algorithme de clustering le plus approprié pour un ensemble de données. Un autre défi est que le nombre de clusters produits peut différer parmi les solutions générées par différents algorithmes de clustering. En outre, le nombre de clusters dans la solution finale est inconnu à l'avance.

[Ghosh et Acharya, 2011] ont souligné qu'il existe plusieurs motivations pour l'utilisation des ensembles de clustering, et que celles-ci sont beaucoup plus larges que celles pour l'utilisation d'un ensemble de classification, où la motivation principale de ce dernier est d'améliorer la précision de la classification. Ces raisons sont les suivantes :

- Améliorer la qualité des résultats du clustering par rapport à ceux produits par des algorithmes de clustering uniques.
- Réutiliser le clustering existant (réutilisation des connaissances) : dans certaines applications, une variété de partitions peut exister, elles peuvent donc être combinées pour obtenir un résultat final de clustering. Ceci permet d'obtenir un résultat de clustering plus consolidé ; plusieurs exemples sont fournis dans [Strehl et Ghosh, 2002].
- Générer des résultats de clustering robustes à travers différents types de jeux de données. Il est largement connu que les algorithmes de clustering populaires échouent souvent à produire un bon résultat de clustering lorsque les données ne correspondent pas à leurs hypothèses.

Parmi ces objectifs, le premier point est le plus largement accepté. La qualité du cluster est généralement mesurée avec une mesure numérique pour évaluer les différents aspects de la validation du cluster [Tan et al., 2006].

Dans cette thèse, on s'est intéressé aux méthodes par ensemble qui consistent à créer un résultat de clustering appelé consensus à partir d'un ensemble de résultats de clustering. Ces méthodes s'intéressent principalement à deux aspects. Le premier est la création des résultats de l'ensemble (différents algorithmes, différentes initialisations, etc.). La seconde est la définition d'une fonction permettant de trouver la partition consensuelle finale.

Contribution

Afin de pallier aux carences des algorihmes individuels de clustering et pour améliorer les résultats, une approche MCS est proposée qui consiste à une application d'un multi- classifieur par l'intégration de l'indice WB sur les données de télédétection. Notre MCS (Multiple Classifier System) est composée de deux étapes essentiels : la génération des partitions et la fonction de consensus.

Pour l'étape de la génération de partitions notre choix s'est porté sur quatre algorithmes de clustering hétérogènes pour concevoir le MCS, à savoir l'algorithme le plus populaire K-means, l'algorithme K-harmonique (KHM), l'algorithme Bisecting K-means (BKM) et l'algorithme neuronale Self Organizing Map (SOM). Le meilleur clustering est obtenu selon un critère d'évaluation de la qualité du clustering à savoir l'indice de validité WB, ce dernier est défini comme le rapport entre la mesure de la compacité du cluster (Sum-of-Squares **W**ithin (SSW)) et sa mesure de séparation (Sum-of-Squares **B**etween (SSB)).

Quant à la fonction de concensus est basée sur le principe de réétiquetage et de vote. Cette approche tente de résoudre le problème de correspondance, puis un simple vote peut être appliqué pour affecter les objets dans les clusters afin de déterminer la partition consensuelle finale.

Structure de la thèse

La présente thèse s'articule autour de quatre chapitres :

Le premier chapitre décrit d'une manière générale la télédétection, les notions, les différents concepts, les domaines d'applications et l'interprétation visuelle, y sont également présentés.

Dans la chaîne de traitement que nous avons élaborée, on est amené à appliquer un processus de classification non supervisé. Nous avons donc jugé utile de donner dans le deuxième chapitre un aperçu général sur les différentes techniques de classification.
Un intérêt particulier a été accordé à quatres algorithmes de clustering à savoir l'algorithme K-means, l'algorithme K-harmonique means (KHM), l'algorithme Bisecting K-means (BKM) et l'algorithme Self Organizing Map (SOM) appliqué dans notre étude.
Enfin, les différents critères d'évaluation de la classification seront présentés dans la dernière partie de ce chapitre.

Le troisième chapitre est dédié aux différentes méthodes d'ensemble de clustering. Nous commençons par aborder l'origine du concept ensemble clustering, ses motivations, ses objectifs et ses applications, pour ensuite présenter les différentes méthodes d'ensemble. Ensuite, nous présentons

l'architecture de notre approche.

Le dernier chapitre constitue un ensemble d'expérimentations de l'approche proposée. Il décrit en détail les tests et la validation du notre MCS sur les différentes expérimentations menées à la fois sur des données artificielles, des images composites, et des données de télédétection.
Une comparaison des résultats obtenus par rapport à ceux donnés par une classification supervisée est présentée à la fin de ce chapitre.

La thèse se termine par une conclusion. Elle donne un aperçu général du travail accompli, met en évidence les résultats obtenus et lance les premiers jalons pour l'extension de ce présent travail.

Télédétection

1

Sommaire

Les données de télédétection sont utilisées dans de nombreux domaines et pour une grande variété d'applications. Ce chapitre aborde donc brièvement les définitions et les concepts de base ainsi que les notions liées à la télédétection.

1.1 Introduction

La télédétection (« Remote sensing » en anglais) est l'art et la science d'enregistrer, de mesurer et d'analyser des informations sur un phénomène à distance. Par trois de nos cinq sens, nous faisons quotidiennement de la télédétection. La vue, l'ouïe et l'odorat nous permettent d'obtenir toute une série d'informations sur notre environnement, de les analyser en temps réel, de les évaluer et de choisir les comportements qui nous semblent les plus adéquats par rapport à la situation observée. Lire le journal, sentir les fleurs sont autant d'activités de télédétection.

La plupart des dispositifs de détection enregistrent des informations sur un objet en mesurant la transmission d'énergie électromagnétique par un objet à partir de surfaces réfléchissantes et rayonnantes. Ces données de télédétection sont largement utilisées dans une série d'applications océanographiques, terrestres et atmosphériques, telles que la cartographie de la couverture terrestre, la modélisation et la surveillance de l'environnement, et la mise à jour des bases de données géographiques.

1.2 Définition de la télédétection

Un certain nombre de définitions différentes, et toutes aussi correctes, de la télédétection sont données ci-dessous :

- La télédétection est une science de l'acquisition, du traitement et de l'interprétation des images qui enregistrent l'interaction entre l'énergie électromagnétique et la matière [Sabins, 1978].
- La télédétection est une instrumentation, des techniques et des méthodes permettant d'observer la surface de la Terre à distance et d'interpréter les images ou les valeurs numériques obtenues afin d'acquérir des informations significatives sur des objets particuliers de la Terre [Buiten et Clevers, 1993].
- La télédétection est définie comme une science et l'art d'obtenir des informations sur un objet, une zone ou un phénomène par l'analyse de données acquises par un capteur qui n'est pas en contact direct avec la cible [Schultz et Engman, 2000].
- La télédétection est une discipline qui cherche à obtenir des informations de la Terre en utilisant des images acquises par des satellites ou des plateformes aériennes et en se servant de la radiation électromagnétique émise ou réfléchie par la surface terrestre [Lillesand et Kiefer, 1994].

1.3 Historique de la télédétection

Le talbeau 1.1 résume quelques dates importantes de l'histoire de la télédétection.

1844	Premières photographies aériennes réalisées depuis un ballon par G.F. Tournachon dit NADAR.
1909	Premières photographies depuis un avion (les frères WRIGHT).
1914 - 1918	Utilisation intensive de la photographie aérienne comme moyen de reconnaissance pendant la 1ère guerre mondiale.
1940	Apparition des premiers radars opérationnels en Grande-Bretagne (bataille d'Angleterre).
1945	Développement continu de la photographie aérienne comme méthode opérationnelle de cartographie et de surveillance de l'environnement.
1957	Lancement de Spoutnik 1, premier satellite artificiel.
1960 - 1972	Développement parallèle de la technique des satellites et des capteurs.
1960	Lancement de Tiros, premier satellite météorologique.
1972	Lancement d'ERTS, rebaptisé Landsat 1, premier satellite spécialisé de télédétection des ressources terrestres.
1982	Apparition de la haute résolution spatiale pour l'observation de la Terre : lancement de Landsat 4, équipé du radiomètre « Thematic Mapper ».
1986	Lancement de SPOT 1 satellite commercial français de télédétection. Développement de capteurs hyperspectraux.
1990-	Développement de systèmes spatiaux à haute résolution.
1999	Lancement par la société privée Space Imaging Corp. du satellite IKONOS, offrant des images à très haute résolution spatiale (1 m).
2001	Lancement de QuickBird, un système de capteurs à très haute résolution spatiale (USA).
2002	Lancement d'Aqua par la NASA. Lancement d'Envisat-1 avec des instruments optiques et radar par l'ESA. Lancement du premier satellite Algérien Alsat 1.
2006	Lancement de ALOS1 (JAPAN).
2008	Lancement de GeoEye.
2009	Lancement de WorldView-2 par DigitalGlobe.
2010	Lancement du deuxième satellite Algérien Alsat 2A.
2013	Lancement de Landsat-8 par la NASA/USGS.
2014	Lancement de WorldView-3 par DigitalGlobe. Lancement de ALOS2 (JAPAN).
2015	Lancement de Sentinel-1 par l'ESA.
2016	Lancement de Sentinel-2 par l'ESA. Lancement de Sentinel-3 par l'ESA. Lancement du troisième satellite Algérien Alsat 2B.
2017	Lancement de Sentinel-5 par l'ESA.
2021	Lancement de Landsat-9 par la NASA/USGS.

Table 1.1 – *Historique de la télédétection.*

Pays opérateur	Nombre de satellites		Pays opérateur	Nombre de satellites
Etats - Unis	887		Espagne	19
Chine	296		Argentine	18
Russie	150		Israel	12
Japan	79		Norvège	8
Multinational	63		Emirats Arabes Unis	8
Inde	57		Singapour	8
ESA (Europe)	49		Kazakhstan	7
Canada	37		Algérie	5
Allamagne	33		Finlande	5
Luxembourg	33		Suisse	3

TABLE 1.2 – *Nombre de satellites opérationnels (2019).*

Selon l'association UCS (Union of Concerned Scientists), 2.787 satellites sont opérationnels au 31 décembre 2020 , dont plus de la moitié lancés par les États-Unis. Les trois quarts des satellites en opération tournent en orbite basse (entre 500 et 2.000 km d'altitude), et sont utilisés pour les systèmes de télécommunication, d'imagerie terrestre ou la météorologie.
Le tableau 1.2 représente le nombre de satellites opérationnels en orbite autour de la Terre de plusieurs nations au 1er avril 2019 [www.futura-sciences.com].

1.4 Processus de télédétection

La télédétection repose sur la mesure de l'énergie électromagnétique ou du rayonnement électromagnétique (EMR). La plus importante source d'énergie à la surface de la Terre est le Soleil, qui nous fournit, par exemple, de la lumière (visible), de la chaleur (que nous pouvons ressentir) et de la lumière UV, qui peut être nocive pour notre peau. De nombreux capteurs utilisés en télédétection mesurent la lumière solaire réfléchie.

La détection et la discrimination d'objets ou de caractéristiques de la surface de la Terre signifie détecter et enregistrer de l'énergie rayonnante réfléchie ou émise par des objets ou des matériaux de la surface (Figure 1.1). Différents objets renvoient une quantité différente d'énergie incidentes dans différentes bandes du spectre électromagnétique. Cela dépend de la propriété du matériau (structurelle, chimique et physique), de la rugosité de la surface, de l'angle d'incidence, de l'intensité et de la longueur d'onde de l'énergie radiante.

La télédétection est fondamentalement une science multidisciplinaire qui comprend une combinaison de diverses disciplines telles que l'optique, la spectroscopie, la photographie, l'informatique, l'électronique, les télécommunications, le lancement de satellites, etc. Toutes ces technologies sont in-

tégrées pour constituer un système complet en soi, connu sous le nom de système de télédétection. Les données de télédétection sont le plus souvent en forme d'images.

Le processus de télédétection comporte un certain nombre d'étapes, chacune d'entre elles étant importante pour le succès de l'opération [Ronald Eastman, 2009].

Les étapes du processus de la télédétection sont :

- Émission de rayonnement électromagnétique (Soleil/auto émission).
- Transmission de l'énergie de la source à la surface de la Terre, ainsi que l'absorption et la diffusion.
- Interaction du rayonnement électromagnétique avec la surface de la Terre : réflexion et émission.
- Transmission de l'énergie de la surface au capteur distant.
- Sortie des données du capteur.
- Transmission, traitement et analyse des données.

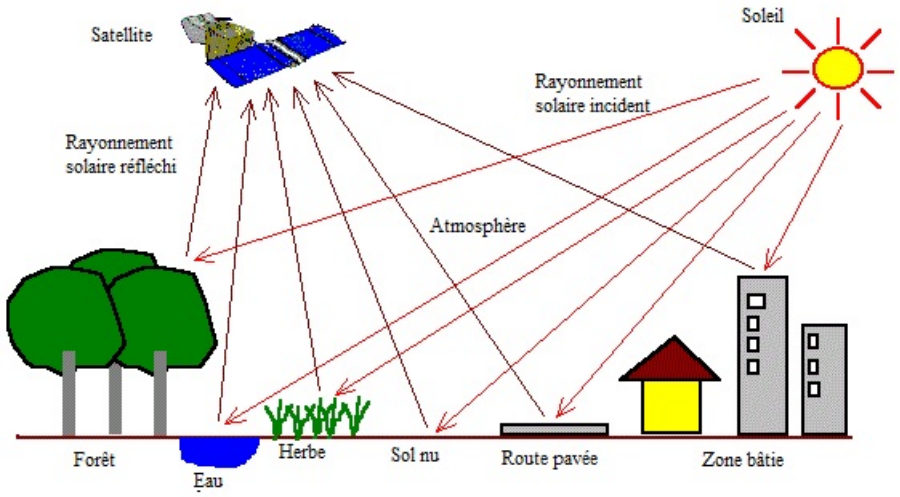


FIGURE 1.1 – *Le système de télédétection.*

La stratégie de base pour détecter le rayonnement électromagnétique est claire. Tout ce qui existe dans la nature a sa propre distribution unique de rayonnement réfléchi, émis et absorbé. Ces caractéristiques spectrales, peuvent être utilisées pour distinguer un objet d'un autre ou pour obtenir des informations sur la forme, la taille et d'autres propriétés physiques et chimiques [Bakker et Al., 2004].

1.4.1 Rayonnement électromagnétique

L'énergie électromagnétique est l'énergie qui se propage sous la forme d'une interaction progressive entre les champs électriques et magnétiques qui sont perpendiculaires l'un à l'autre [Sabins, 1978] (Figure 1.2). Elle se déplace à la vitesse de la lumière. La lumière visible, les rayons ultraviolets, les rayons infrarouges, la chaleur, les ondes radio et les rayons X sont tous

des formes différentes d'énergie électromagnétique.

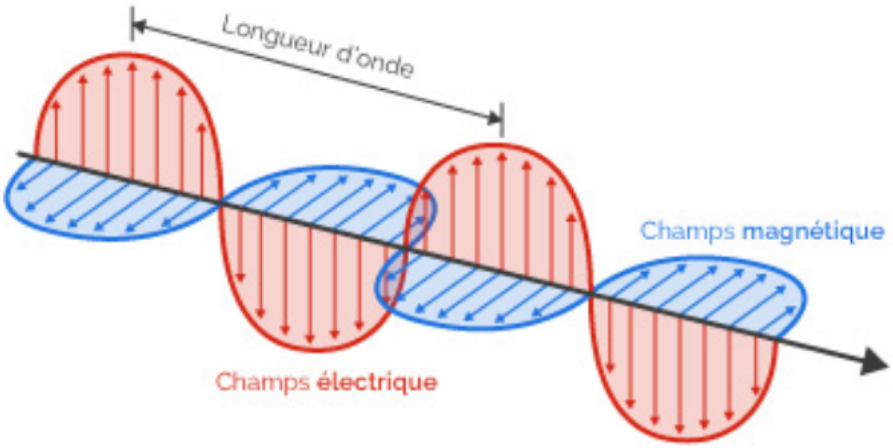


FIGURE 1.2 – *Représentation d'une onde électromagnétique.*

L'énergie électromagnétique (E) peut être exprimée en termes de fréquence (f) ou de longueur d'onde (λ) du rayonnement comme suit :

$$E = h\,c\,f \qquad ou \qquad \frac{h\,c}{\lambda} \tag{1.1}$$

où h est la constante de Planck (6,626 x 10^{-34} Joules-sec), c est une constante qui exprime la vitesse de la lumière (3 x 10^{8} m/sec), f est la fréquence exprimée en Hertz et λ est la longueur d'onde exprimée en micromètres (1μm = 10^{-6} m).

Comme on peut le constater à partir de l'équation (1.1), les longueurs d'onde plus courtes ont un contenu énergétique plus élevé et les longueurs d'onde plus longues ont un contenu énergétique plus faible.

La distribution du continuum d'énergie peut être tracée en fonction de la longueur d'onde (ou de la fréquence) et est connue sous le nom de spectre du rayonnement électromagnétique (Figure 1.3).

Dans la terminologie de la télédétection, l'énergie électromagnétique est généralement exprimée en termes de longueur d'onde (λ).

Tous les matériaux reflètent, émettent ou rayonnent une gamme d'énergie électromagnétique, selon les caractéristiques du matériau. En télédétection, c'est la mesure du rayonnement électromagnétique réfléchi ou émis par un objet, qui est utilisée pour identifier la cible et déduire ses propriétés [Bakker et Al., 2004].

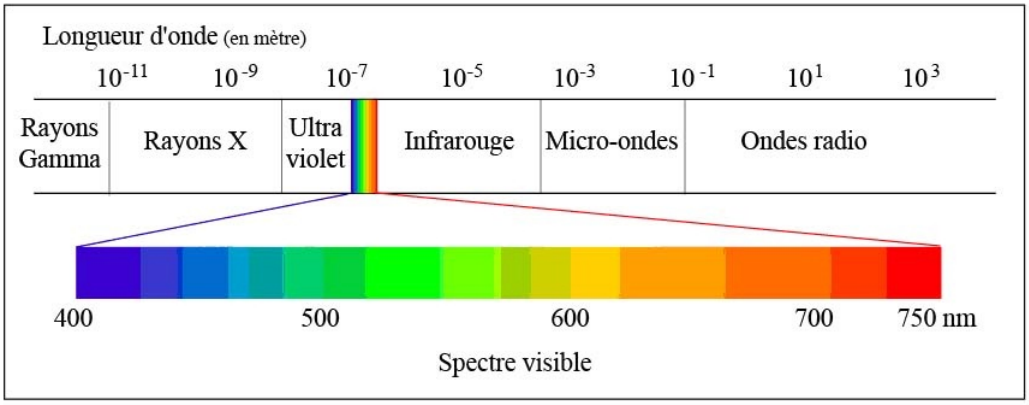


FIGURE 1.3 – *Spectre de rayonnement électromagnétique.*

En télédétection, les longueurs d'ondes les plus utilisées sont :
Le visible 0.4 - 0.7 μm.
Le proche infrarouge 0.7 - 1.5 μm.
L'infrarouge moyen 1.5 – 3 μm.
L'infrarouge thermique 3 - 15 μm.
Les micro-ondes ou hyperfréquences 1 mm – 1 m.

1.4.2 Interaction du rayonnement électromagnétique avec l'atmosphère et la surface terrestre

Comme vous le savez, le mot "atmosphère" désigne les couches de gaz qui entourent la Terre. Les constituants de l'atmosphère sont l'azote, l'oxygène, le dioxyde de carbone, l'ozone, la vapeur d'eau et d'autres gaz. Le rayonnement électromagnétique provenant du Soleil doit traverser deux fois l'atmosphère terrestre avant d'être détecté par le capteur du satellite, une fois lors de son voyage du Soleil vers la Terre et une seconde fois après avoir été réfléchi/émis par la surface de la Terre vers le capteur. Les particules et les gaz présents dans l'atmosphère interagissent avec la lumière entrante et le rayonnement réfléchi/émis. L'intérêt que nous portons à cette interaction est lié au fait que les composants atmosphériques diffusent, réfractent, réfléchissent, absorbent et émettent des rayonnements électromagnétiques, modifiant ainsi la radiance originale des objets observés par un capteur à distance [Bakker et Al., 2004].

L'interaction du rayonnement électromagnétique avec l'atmosphère est importante pour la télédétection pour deux raisons principales :

- L'information transportée par le rayonnement électromagnétique réfléchi/émis par la surface de la Terre est modifiée lorsqu'elle traverse l'atmosphère.
- L'interaction du rayonnement électromagnétique avec l'atmosphère peut être utilisée pour obtenir des informations utiles sur l'atmosphère elle-même.

1.5 Télédétection passive/active

Selon la source d'énergie électromagnétique, la télédétection peut être classée en télédétection passive ou active (Figure 1.4).

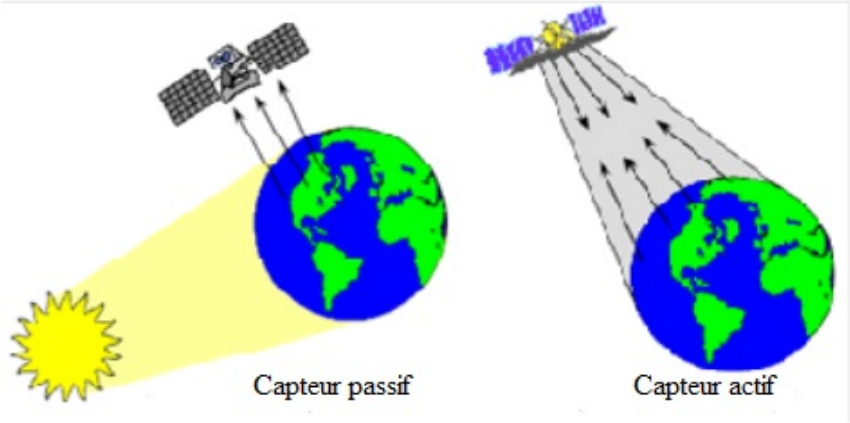


Figure 1.4 – *Télédétection passive et active.*

Dans le cas de la télédétection passive, la source d'énergie est celle qui est naturellement disponible, comme le Soleil. La plupart des systèmes de télédétection fonctionnent en mode passif en utilisant l'énergie solaire comme source de rayonnement électromagnétique. L'énergie solaire réfléchie par les cibles à des bandes de longueur d'onde spécifiques est enregistrée à l'aide de capteurs embarqués sur des plates-formes aériennes ou spatiales. Afin de garantir une puissance de signal suffisante au niveau du capteur, des bandes de longueur d'onde/énergie capables de traverser l'atmosphère, sans perte significative due aux interactions atmosphériques, sont généralement utilisées en télédétection.

Tout objet dont la température est supérieure à 0^0 K (Kelvin) émet un rayonnement, qui est approximativement proportionnel à la puissance quatre de la température de l'objet. Ainsi, la Terre émet également un certain rayonnement puisque sa température ambiante est d'environ 300^0 K.

Dans le cas de la télédétection active, l'énergie est générée et envoyée de la plate-forme de télédétection vers les cibles. L'énergie renvoyée par les cibles est enregistrée à l'aide de capteurs embarqués sur la plate-forme de télédétection. La plupart de la télédétection par micro-ondes est effectuée par la télédétection active [Bakker et Al., 2004].

Donc, selon les cas nous aurons les informations présentées dans le tableau 1.3.

Par simple analogie, la télédétection passive est similaire à la prise de vue avec un appareil photo ordinaire, tandis que la télédétection active est analogue à la prise de vue avec un appareil photo doté d'un flash intégré.

Longueur d'onde utlisée	300 - 1100 nm	8000 à 14000	>14000 nm
Domaine du spectre électro - magnétique	visible et proche infra -rouge	infra-rouge thermique	infra-rouge lointain micro - ondes
Type de télédétection	Passive	Passive	Active
Conditions	Diurne	Diurne et nocturne	Diurne et nocturne

TABLE 1.3 – *Domaines du spectre électromagnétique de la télédétection passive et active.*

1.6 VECTEURS

Par capteur nous désignons l'appareil qui enregistre le rayonnement électromagnétique venant du sol, et par vecteur (plate-forme) l'avion, le ballon ou le satellite qui le transporte.

Les plates-formes spatiales se trouvent dans l'espace et se déplacent sur leur orbite autour de la Terre. C'est grâce à ces plates-formes spatiales que nous obtenons d'énormes quantités de données de télédétection.

En fonction de leur altitude et de leur orbite, ces plateformes peuvent être divisées en deux catégories : les satellites géostationnaires et les satellites à défilement (Figure 1.5).

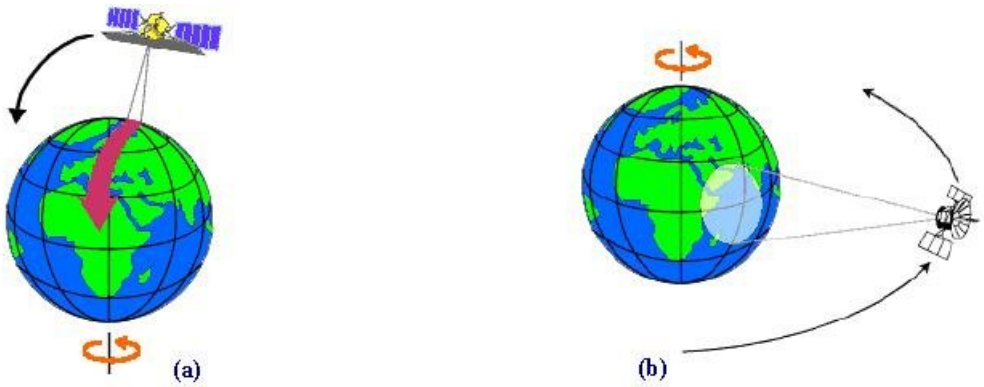


FIGURE 1.5 – *Types de plates-formes*
a) Orbite polaire. b) Orbite géostationnaire.

1.6.1 Satellites géostationnaires

Un satellite géostationnaire est un satellite artificiel qui se trouve sur une orbite géostationnaire.

Un satellite sur cette orbite située à 35 786 km d'altitude possède une période de révolution très exactement égale à la période de rotation de la Terre et paraît immobile par rapport à un point de référence à la surface de la Terre, c'est-à-dire reste toujours à la verticale du même point sur Terre, propriété utilisée pour en faire des satellites d'observation, de télécommunications, ou bien de télédiffusion [Bakker et Al., 2004]. Pour respecter cette propriété, un satellite géostationnaire se situe forcément dans le plan de l'équateur. Son taux de rotation – sa vitesse angulaire de rotation – est égal à celui de la Terre, soit environ 15^0/heure.

Les satellites METEOSAT (France), GOES (Geostationary Operational Environmental Satellites, USA), GMS (Japon), INSAT (Inde) et le satellite Algérien de télécommunication ALCOMSAT-1 sont géostationnaires.

1.6.2 Satellites à défilement

Les satellites à défilement gravitent sur une orbite polaire quasi circulaire à une altitude de 800 à 1000 kilomètres. La rotation s'effectue en 100 minutes, ce qui équivaut à un passage au dessus d'un même point terrestre toutes les 12 heures environ.

Grâce à ces satellites, le globe entier est couvert de façon régulière et donne une couverture répétitive sur une base périodique [Bakker et Al., 2004].

Les satellites à défilement sont principalement utilisés pour la météorologie, pour les systèmes de téléphonie planétaire, le GPS (Global Positioning System, en anglais) qui est le principal système de positionnement satellite mondial actuellement et qui orbite aux alentours de 20000 km d'altitude, ou encore les stations spatiales comme Mir (Russie) ou, actuellement, ISS (Station Spatiale Internationale, International Space Station, en anglais).

Les satellites français SPOT, le programme américain LANDSAT et les satellites Algérien Alsat circulent sur ce type d'orbite.

1.7 Avantages et inconvénients de la télédétection

Les avantages de la télédétection sont les suivants :

- Fournir des données sur des grandes surfaces.
- Fournir des données sur des régions très éloignées et inaccessibles.
- Capable d'obtenir des images de n'importe quelle zone sur une période de temps continue, ce qui permet d'analyser les changements anthropiques ou naturels du paysage.
- Relativement peu coûteuse par rapport à l'emploi d'une équipe de géomètres.
- Collecte facile et rapide de données.
- Production rapide de cartes pour l'interprétation.

Ses inconvénients sont :

- L'interprétation de l'imagerie requiert un certain niveau de compétence.
- Nécessité d'une vérité terrain.
- Les données provenant de sources multiples peuvent créer de la confusion.
- Les objets peuvent être mal classifiés ou confondus.
- Des distorsions peuvent se produire dans une image en raison du mouvement relatif du capteur et de la source.

1.8 Données de télédétection

1.8.1 Images satellitaires

L'imagerie satellitaire (aussi appelée imagerie spatiale) désigne la prise d'images depuis l'espace, par des capteurs placés sur des satellites.

Les images satellitaires peuvent être classées selon le mode d'acquisition (images panchromatiques, images multi-spectrales, images hyper-spectrales).

Les images panchromatiques

Une bande panchromatique (bande noire et blanche) est une bande qui contient généralement une largeur de bande de quelques centaines de nanomètres (entre 0.4 et 0.7 μm). Cette largeur de bande lui permet de conserver un signal-bruit élevé, ce qui rend les données panchromatiques disponibles à une haute résolution spatiale et une faible résolution spectrale (Figure 1.6).

Figure 1.6 – *Image panchromatique du satellite Quickbird de résolution 61 cm d'une région au Canada.*

Les images multi-spectrales

L'imagerie multispectrale fait généralement référence à des enregistrements simultanés dans un petit nombre de bandes spectrales (3 à 10 bandes).

La figure 1.7 donne une schématisation d'une image satellitaires qui se représente sous la forme d'une matrice de pixels. Chaque pixel est décrit par un ensemble de valeurs correspondant aux bandes du capteur utilisé pour capturer l'image.

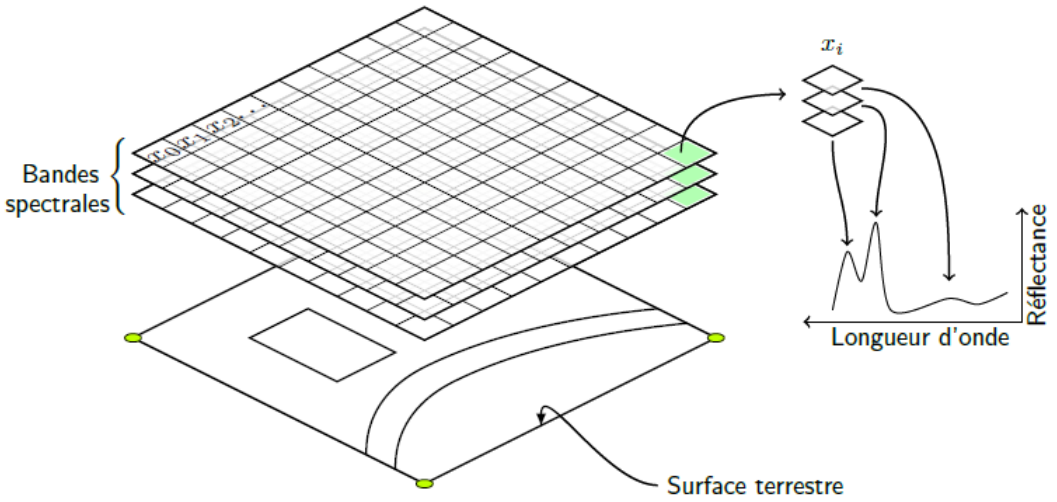


FIGURE 1.7 – *Illustration d'une image satellitaires multispectrale.*

Les images hyper-spectrales

Les images hyperspectrales sont obtenues par des capteurs capables d'enregistrer l'information dans une multitude (souvent plus de 200) de bandes spectrales beaucoup plus étroites (de l'ordre de quelques nm) et souvent contigües, dans les portions visible, proche infrarouge et infrarouge moyen du spectre électromagnétique (Figure 1.8). Par exemple, le satellite américain Hypérion compte 220 bandes spectrales (à partir de 0.4 à 2.5 mm) avec 30 m de résolution spatiale.

Chaque matériau possède une signature spectrale spécifique qui peut être utilisée comme une "empreinte digitale" pour son identification unique. Par conséquent, l'imagerie hyperspectrale trouve un large éventail d'applications dans la télédétection tels que l'astronomie, l'agriculture, la biologie moléculaire, l'imagerie biomédicale, la minéralogie, la géologie, la physique, le patrimoine culturel, l'agroalimentaire, l'environnement et la surveillance.

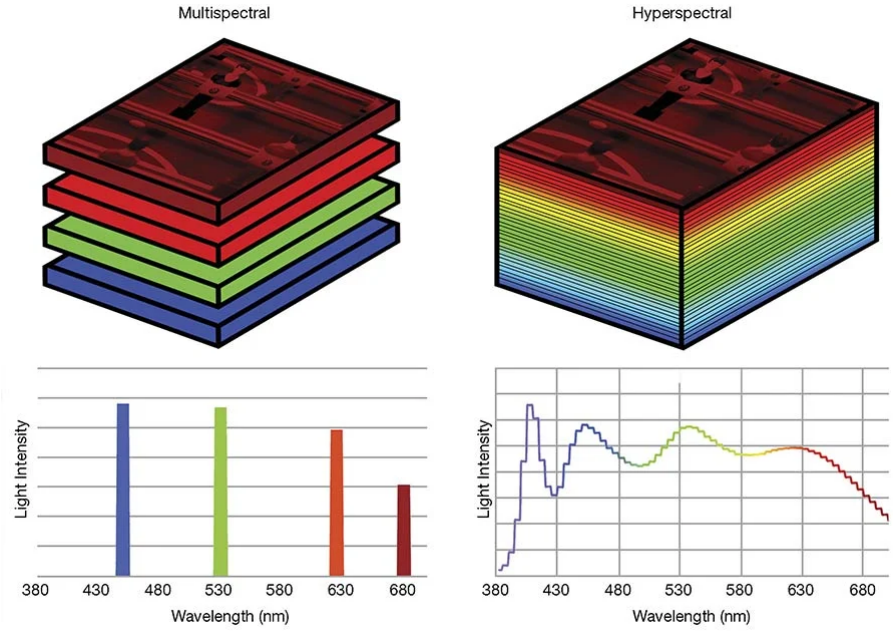


FIGURE 1.8 – *Comparaison d'une série d'images en imagerie multispectrale, dans laquelle des images sont prises dans plusieurs spectres différents, et en imagerie hyperspectrale, dans laquelle des images sont prises dans de nombreux spectres différents [Bodersen, 2021].*

1.8.2 Résolution

En télédétection, le terme résolution est utilisé pour représenter le pouvoir de résolution, qui comprend non seulement la capacité d'identifier la présence de deux objets, mais aussi leurs propriétés. En termes qualitatifs, la résolution est la quantité de détails qui peuvent être observés dans une image [Lillesand et Kiefer, 1994].

Quatre types de résolutions sont définis pour les systèmes de télédétection : la résolution spatiale, spectrale, radiométrique et temporelle.

Résolution spatiale

La résolution spatiale mesure la plus petite séparation angulaire entre deux objets. Dans le cas des images satellitaires, elle est représentée en pixels et la résolution spatiale d'une image donnée est notée comme le nombre de mètres que représente ce pixel. Par exemple, le satellite multispectral SPOT 4 a une résolution spatiale de 20m. Cela signifie que chaque pixel carré individuel représente une zone spatiale de 400 mètres carrés. Dans le domaine civil, on distingue généralement, la basse résolution (BR) (1000m), la moyenne résolution (MR) (80m), la haute résolution (HR) (10 à 30m) et la très haute résolution (THR) (inférieure à 5m) (Figure 1.9).

FIGURE 1.9 – *Image satellitaire Quickbird de résolution 2,44 m d'une région au Canada.*

Résolution spectrale

La résolution spetrale désigne la capacité d'un capteur satellitaire à mesurer des longueurs d'onde spécifiques du spectre électromagnétique.

- Plus la résolution spectrale est fine, plus la gamme de longueurs d'onde pour un canal ou une bande particulière est étroite.
- Plus la résolution spectrale est élevée, plus il est possible d'exploiter les différences entre les signatures spectrales.

Résolution radiométrique

La résolution radiométrique représente la sensibilité du capteur à la magnitude de l'énergie électromagnétique. Plus la résolution radiométrique d'un capteur est fine, plus il est sensible à la détection de petites différences dans l'énergie réfléchie ou émise ou, en d'autres termes, le système peut mesurer un plus grand nombre de niveaux de gris.

Résolution temporelle

La résolution temporelle est le temps pris par le capteur à bord du satellite pour capturer des images successives du même endroit sur la surface de la Terre. En d'autres termes, la résolution temporelle est le temps de revisite

ou le cycle de répétition du satellite sur la même région ou le même endroit à la surface de la Terre (par exemple, la résolution temporelle de la série IRS (Inde) est de 24 jours, celle de la série SPOT (France) de 26 jours, celle d'IKONOS (USA) de 2,9 jours, etc.).

On trouve aussi la constellation de satellite à savoir un ensemble de satellites remplissant des fonctions identiques, répartis de façon à assurer, en permanence une couverture quasi-complète de la planète, par exemple les deux satellites jumeaux, Sentinel-2A et Sentinel-2B (ESA).

1.9 Analyse et interprétation des images

L'interprétation d'images est une technique puissante qui nous permet d'identifier et de distinguer diverses caractéristiques dans les images de télédétection et d'acquérir des connaissances et des informations à leur sujet. L'analyse d'une image de télédétection implique souvent l'identification de diverses caractéristiques telles que la couverture forestière, les plans d'eau, l'habitat urbain, et l'agriculture, etc. Ces caractéristiques sont identifiées par la façon dont elles reflètent ou émettent des rayonnements électromagnétiques et aussi par leur association et leur emplacement. Ces rayonnements électromagnétiques sont mesurées par des capteurs satellites et sont finalement représentées sous forme d'image satellite [Sabins, 1978].

Parmi les différents techniques d'interprétation on a :

Reconnaissance : c'est une tentative de distinguer un objet par ses caractéristiques ou ses motifs sur l'image.

Analyse : c'est un processus de résolution ou de séparation d'un ensemble d'objets ou d'éléments ayant des caractéristiques similaires.

Classification : est définie comme le processus de catégorisation de tous les pixels d'une image pour obtenir un ensemble d'étiquettes ou de thèmes d'occupation du sol.

1.9.1 Éléments d'interprétation

L'interprétation des images satellites implique l'étude de divers caractères de base d'un objet en référence aux bandes spectrales, ce qui est utile dans l'analyse visuelle. Les éléments de base sont la forme, la taille, le patron, le ton, la texture, les ombres et l'association [Lillesand et Kiefer, 1994]. L'identification des cibles en télédétection basée sur les 7 caractéristiques visuelles nous permet de mieux interpréter et analyser.

- **La forme** réfère à l'allure générale, la structure ou le contour des objets. Les formes aux bordures rectilignes se retrouvent généralement dans les régions urbaines ou dans des champs agricoles, alors que les structures naturelles, telles que les bordures des forêts, sont généralement plus irrégulières.

- **La taille** dépend de l'échelle et de la résolution de l'image. Il est important d'évaluer la taille d'une cible par rapport aux autres objets dans une scène (taille relative), ainsi que la taille absolue, afin d'aider l'interprétation de cette cible.

- **Le patron** réfère à l'agencement spatial des objets visiblement discernables. Une répétition ordonnée de tons similaires et de textures produit un patron distinctif et facilement reconnaissable. Les vergers avec leurs arbres régulièrement disposés, ou les rues régulièrement bordées de maisons sont de bons exemples de patrons.

- **Le ton** réfère à la clarté relative ou la couleur (teinte) des objets dans une image. Généralement, la nuance de ton est l'élément fondamental pour différencier les cibles et les structures. Les variations de ton permettent aussi la différenciation des formes, textures et patrons des objets.

- **La texture** réfère à l'arrangement et à la fréquence des variations de teintes dans des régions particulières d'une image. Des textures rugueuses consisteraient en des tons en rayures où les niveaux de gris changent brusquement dans une petite région, alors que les textures lisses auraient peu ou pas de variations de tons. Les textures lisses sont souvent le résultat de surfaces uniformes telles que des champs, du pavement ou des terrains gazonnés. Une cible avec une surface rugueuse et une structure irrégulière, telle qu'une forêt, résulte en une texture d'apparence rugueuse.

- **L'ombre** indique le contour d'un objet et sa longueur, ce qui est utile pour mesurer la hauteur d'un objet. En imagerie radar, les ombres sont particulièrement utiles pour rehausser ou identifier la topographie et les formes géologiques.

- **L'association** tient compte de la relation entre la cible d'intérêt et d'autres objets voisins. L'identification d'éléments qu'on s'attend normalement à retrouver à proximité d'autres structures peut donner de l'information facilitant l'identification.

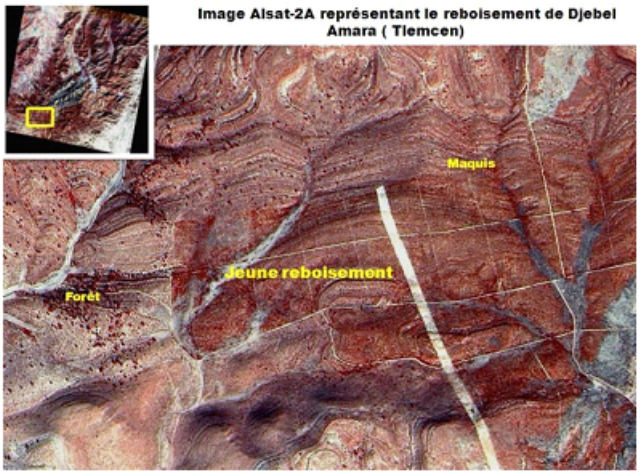
1.10 Domaines d'application de la télédétection

Les applications de la télédétection se sont multipliées, dans de nombreux domaines tels que :

- **La météorologie** (Evolution spatio-temporelle de la couverture nuageuse, mesures de température, de vapeur d'eau et de précipitations...).
- **L'agriculture** (Prévisions de récoltes, évaluation des dommages causés par la sécheresse et les inondations,...).
- **La foresterie** (Densité des forêts, estimation de la biomasse, déforestation, suivi et évaluation des feux,...).
- **L'urbanisme** (Estimation de la population, planification des ports, aéroports, routes...).
- **La géologie** (Exploration minière, exploration pétrolière, cartographie et surveillance des taux de sédimentation, cartographie et surveillance des phénomènes naturels...).
- **l'hydrologie** (Cartographie et surveillance des marécages, évaluation de l'humidité du sol, cartographie et la modélisation des bassins hydrologiques, ...).
- **La topographie** (Modèles numériques d'élévation de terrain).
- **La glaciologie** (Surveillance des glaciers, des icebergs).
- **La défence**.
- **L'archéologie**.
- **La géodésie**.
- **L'océanographie**.

les figures 1.10 et 1.11 montrent des exemples des différents applications spatiales du satellite algérien Alsat 2A.

La limite entre la forêt incendiée (teinte grise) et celle non incendiée (teinte rouge), représentée par une tranchée pare-feu.



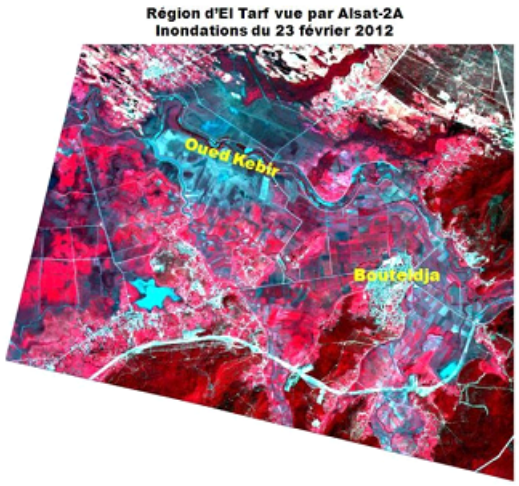
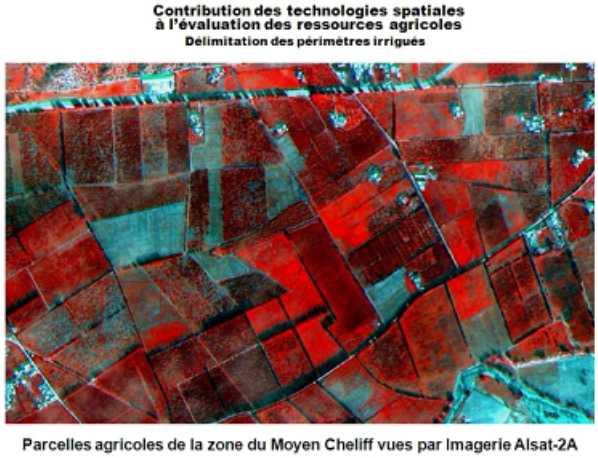


FIGURE 1.10 – *Applications spatiales du satellite Alsat-2A (a) [www.asal.dz].*



Parcelles agricoles de la zone du Moyen Cheliff vues par Imagerie Alsat-2A

FIGURE 1.11 – *Applications spatiales du satellite Alsat-2A (b) [www.asal.dz].*

1.11 Différents systèmes de télédétection

Actuellement, un certain nombre de grands systèmes de satellites de ressources terrestres tournent autour du monde.

Le tableau 1.4 résume les caractéristiques des plates-formes, des orbites et des capteurs des principaux systèmes de satellites dans le monde.

Satellite	**Aster**	**Ikonos**	**Spot**	**Landsat**	**Alsat 2A**	**Radarsat**
Pays	USA	USA	France	USA	Algérie	Canada
Type	hélio-synchrone	hélio-synchrone	hélio-synchrone	hélio-synchrone	hélio-synchrone	hélio-synchrone
Altitude	705 km	synchrone	832 km	920 km * 705 km **	672 km	798 km
Inclinaison (degré)	98,3	98,1	98,7		98	98,6
Période (min)	98,88		101,4	100		100
Cycle de répétition (jours)	16	14	26	18 * 16 **	3	24
Rés. spatiale (mètre)	15 30 90	1 PAN 4 MS	10 PAN 20 MS	15 PAN 30 TM 80 MSS	2,5 PAN 10 MS	8 - 100

Table 1.4 – *Caractéristiques de certains satellites.*

* Landsat 1, 2, 3.
** Landsat 4, 5.

1.11.1 Programme ALSAT

Le programme Alsat est une famille de satellites artificiels algériens gérés par l'Agence Spatiale ALgérienne (ASAL). L'exploitation des données images issues des satellites algériens permet le renforcement de capacité de prise de vue, répond aux besoins et aux préoccupations des différents secteurs utilisateurs, notamment la révision cartographique, le cadastre, l'environnement et l'aménagement du territoire, les ressources naturelles, l'urbanisme et les infrastructures de base [www.asal.dz].

Alsat 1

C'est le premier satellite artificiel algérien. Alsat-1 est un micro-satellite de cartographie lancé le 28 novembre 2002. Il a pour objectif principal la gestion des ressources naturelles du territoire algérien. Son orbite est héliosynchrone et il a été placé à une altitude approximative de 700 kilomètres sur inclinaison de 98°. Le système d'imagerie couvre le vert, le rouge et le proche infrarouge, pour une résolution de 32 mètres (Figure 1.12).

FIGURE 1.12 – *Image Alsat1 (32m) de la région d'Oran.*

Alsat-2A

Alsat-2A est le deuxième satellite d'observation de la Terre du programme spatial national mis en orbite après Alsat-1, lancé le 12 juillet 2010. C'est un satellite d'observation de la Terre à haute résolution, avec une résolution spatiale de 2.5 m en mode panchromatique, et de 10 m en mode multispectral couvert par les 04 bandes spectrales suivantes : le bleu, le vert, le rouge et le proche infrarouge, avec une répétitivité de 03 jours et un champ d'observation de 17,5 Km.

Alsat-2B

Alsat-2B, troisième satellite d'observation de la Terre a été mis en orbite le 26 septembre 2016, son orbite opérationnelle a une altitude de 670 Km avec un déphasage de 186 degré par rapport à son jumeau Alsat-2A. Les deux satellites forment une mini-constellation, ce qui permet d'augmenter la fréquence de prise de vues .

L'exploitation des données images issues des deux satellites Alsat-2A et Alsat-2B permet le renforcement de capacité de prise de vue, répond aux besoins et aux préoccupations des différents secteurs utilisateurs, notamment la révision cartographique, le cadastre, l'environnement et l'aménagement du territoire et les ressources naturelles.

Alsat-1B

Alsat-1B, le quatrième satellite d'observation de la Terre, il assure la continuité de la couverture nationale antérieurement assurée par le satellite Alsat-1 ayant fournit depuis 2002 des images à moyenne résolution.

Conclusion

Dans ce chapitre, nous avons abordé les définitions et les concepts de base ainsi que les notions liées à la télédétection. Les données de télédétection contiennent une grande quantité d'informations. Il existe de nombreuses méthodes et algorithmes disponibles pour le traitement des images satellites tels que le rehaussement, l'extraction de caractéristiques, la segmentation, la fusion, la détection de changements, la compression, la classification et la détection de caractéristiques.

Parmis toutes ces méthodes, on s'intéresse, dans ce qui suit, à la classification qui sera utilisée le long de ce travail.

Classification des images satellitaires

2

Sommaire

La classification est une étape essentielle du processus de vision. Elle a pour objectif de partitionner l'image en zones stationnaires qu'on espère les plus proches de la réalité. Dans ce chapitre, nous présenterons brièvement les méthodes de classification supervisées. Puis, nous exposerons par la suite les méthodes non supervisées, ainsi que les algorithmes de clustering appliqués à notre étude. Les critères d'évaluation de la classification seront présentés dans la dernière partie de ce chapitre.

2.1 Introduction

L'apprentissage automatique en télédétection comprend plusieurs méthodes différentes tels que la classification, le clustering, la régression, l'extraction de caractéristiques, la réduction de la dimensionnalité et l'estimation de la densité.

Ces aspects sont souvent interdépendants, par exemple, avant d'effectuer une classification, on peut extraire des caractéristiques de texture supplémentaires et réduire la dimensionnalité de l'ensemble de données à l'aide de techniques de sélection de caractéristiques (Figure 2.1). Les applications les plus courantes en télédétection sont sans doute la réduction des caractéristiques, le clustering et la classification.

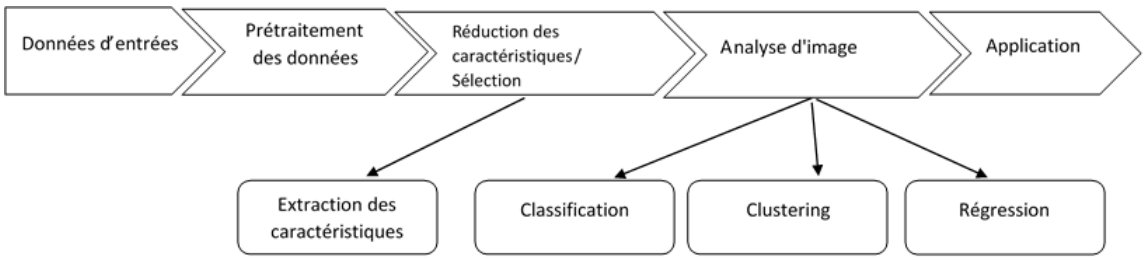


Figure 2.1 – *Aperçu d'une application informatique en télédétection [Camps-Valls et Bruzzone, 2009].*

La classification d'images fait partie des méthodes couramment utilisées pour exploiter des images satellitaires d'observation de la terre [Mather, 2009]. Elle consiste à affecter à chaque entité de la scène traitée une étiquette indiquant son appartenance à une classe particulière. L'entité utilisée caractérise généralement le pixel ou une région de pixels tandis que l'étiquette constitue un thème choisi par l'utilisateur, par exemple la végétation, pour quantifier l'occupation des sols.

Le présent chapitre est donc consacré à l'exposé des méthodes de classification.

Les méthodes ne prenant en compte qu'un seul attribut (le niveau de gris) sont qualifiées de méthodes monodimensionnelles. Ces méthodes se basent sur la détection des seuils à partir de l'histogramme multimodal. Chaque seuil représente ainsi une classe. L'affectation d'un pixel à une classe se fait par comparaison de la valeur de pixel par rapport aux différents seuils.

Les méthodes exploitant plusieurs attributs sont qualifiées de multidimensionnelles. Elles utilisent des algorithmes d'analyse de données.

Le type d'une méthode de classification multidimensionnelle se décline généralement en trois familles : le mode supervisé, le mode non supervisé et le mode semi-supervisé.

Si l'on dispose d'un ensemble de points étiquetés, on parlera de classification supervisée. Dans le cas contraire, on effectue une classification non supervisée (Clustering en anglais).

Par contre, l'apprentissage semi-supervisé est une approche de l'apprentissage automatique qui combine une petite quantité de données étiquetées avec une grande quantité de données non étiquetées pendant la formation. L'apprentissage semi-supervisé se situe entre l'apprentissage non supervisé (sans données de formation étiquetées) et l'apprentissage supervisé (avec uniquement des données de formation étiquetées).

Dans cette partie, nous exposons les différentes méthodes de classification et nous mettons l'accent sur la classifciation non supervisée.

2.2 Méthodes monodimensionnelles ou seuillage

L'attribut analysé est le niveau de gris. Une analyse de l'histogramme de l'image permet de dégager automatiquement des groupes et donc des seuils les séparant [Cocquerez et Philipp, 1995][Duchesnay, 2001].

Une image, à titre d'exemple, représentant des objets foncés sur un fond clair présentera un histogramme avec deux modes bien distincts (Figure 2.2).

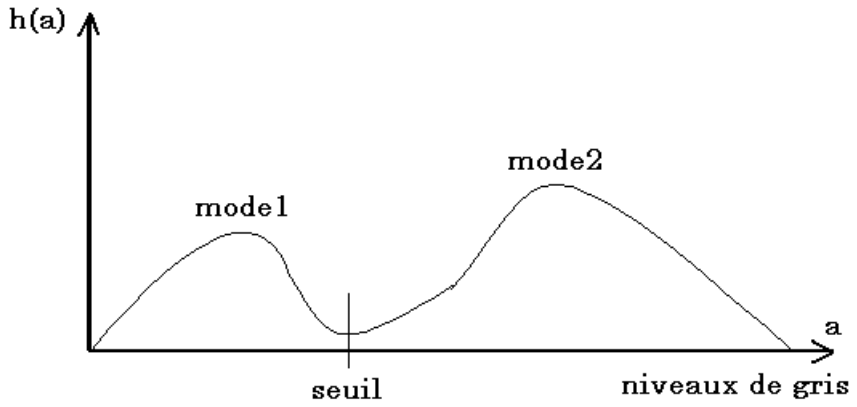


Figure 2.2 – *Histogramme bimodal illustrant la présence de deux classes d'intensité dans l'image.*

D'une manière précise, la méthode de classification monodimensionnelle se réalise en trois étapes :

- Identification des seuils interclasses.
- Affectation des points aux différentes classes.
- Extraction des composantes connexes de chaque classe.

2.3 Méthodes multidimensionnelles

Les approches multidimensionnelles consistent à classifier des individus en fonction non plus d'un seul attribut (niveau de gris), mais en fonction d'un ensemble d'attributs. Elles permettent de traiter les images multi-spectrales dont la simple analyse du niveau de gris se révèle insuffisante.

On distingue deux types de méthodes de classification multidimensionnelles. La première, appelée classification supervisée (dirigée), et la seconde appelée classification non supervisée (non dirigée) (Figure 2.3).

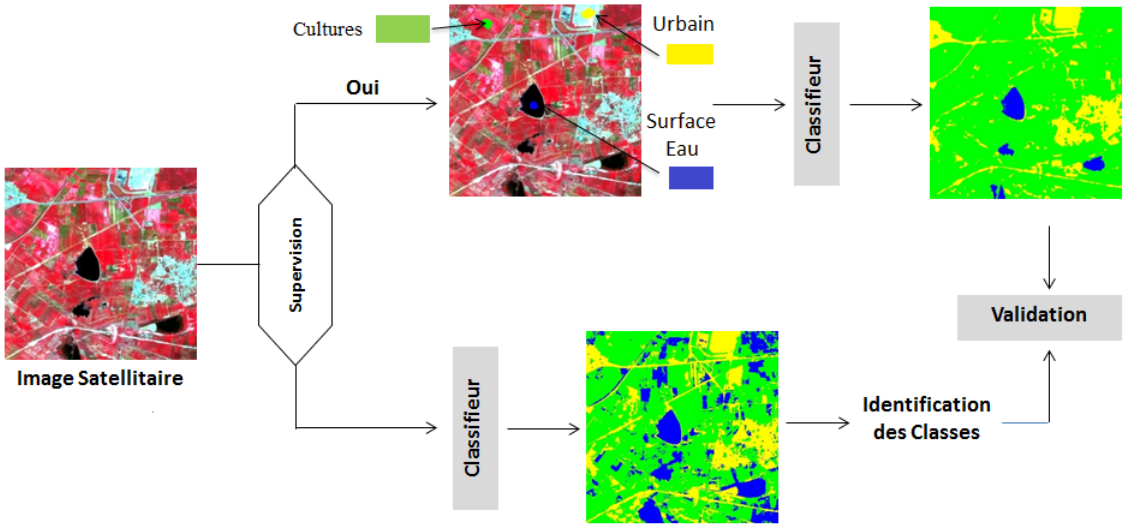


Figure 2.3 – *Types de classification.*

2.4 Classification supervisée

La Classification supervisée ou classification avec apprentissage, est l'une des méthodes les plus utilisées en classification des images de la télédétection. Cette méthode suppose une connaissance à priori des classes à obtenir, du fait d'une connaissance préalable du terrain.

Plusieurs algorithmes de classification supervisée ont été développés, partant de ceux basés sur les modèles probabilistes des classes à distinguer jusqu'à ceux qui partitionnent l'espace des caractéristiques en des régions spécifiques représentant les classes, en utilisant des critères géométriques.

Quelque soit l'algorithme utilisé, les étapes pratiques d'une classification supervisée sont :

1. Choisir les classes d'objets (thèmes) à identifier dans l'image ;
2. Sélectionner les données d'apprentissage (échantillons) représentatives des classes désirées, en utilisant des cartes thématiques, des images classifiées de la même scène ou à partir d'étude sur le terrain ;
3. Utiliser les données d'apprentissage pour estimer les paramètres du classificateur choisi. Ces paramètres peuvent être ceux des modèles probabilistes adoptés pour les classes, où ceux des équations qui définissent les partitions de l'espace de mesure ;
4. Classifier l'image entière ;
5. Evaluer l'exactitude de la classification en calculant la matrice de confusion.

Les méthodes supervisées peuvent être divisées en deux groupes :
- Celles utilisant des hypothèses probabilistes (statistiques) ou parammétriques,
- Celles utilisant des hypothèses géométriques ou non paramétriques.

2.4.1 Méthodes probabilistes (statistiques) ou paramétriques

Pour ces méthodes on pose souvent l'hypothèse gaussienne [1](moyenne, écart type, variance, covariance).

Afin d'avoir une idée sur le rassemblement des nuages de points, on effectue un examen des éléments statistiques. Pour cela, nous avons divers types de traitements pour pouvoir discriminer ces classes.

Parmi les méthodes probabilistes les plus utilisées, sont :

Basée sur une logique booléenne, **la méthode du parallélépipède** prend en considération les statistiques du premier ordre (moyenne et variance) calculées à partir d'échantillons d'apprentissage. Les seuils sont fixés implicitement par le maximum et le minimum des valeurs de réflectance pour chaque classe et pour chaque bande spectrale. Ainsi les bornes supérieures et inférieures de chaque classe délimitent des régions à l'intérieur de l'espace tridimensionnel (Cas d'une image à trois composantes) des données et forment ainsi des règles de décision. Ces régions prennent la forme d'une boite avec des cotés parallèles et perpendiculaires donnant la forme d'un parallélépipède, d'où le nom de l'algorithme. Si la valeur du pixel est comprise entre la borne inférieure et la borne supérieure pour les trois canaux pour une classe donnée, il est affecté à cette classe, dans le cas contraire il est affecté à une catégorie non classée. Dans le cas, où un pixel vérifie toutes les conditions pour plusieurs classes à la fois (cas de chevauchement de certains

1. On peut aussi trouver d'autres méthodes s'appuyant sur une hypothèse de distribution autre que la gaussienne.

parallélépipèdes), il est affecté automatiquement à la première classe.

Le principe de **la méthode de distance minimum** consiste à chercher la classe la plus proche pour chaque pixel, ou groupe de pixels si l'on travaille dans une fenêtre d'analyse centrée sur le pixel courant. La notion de proximité est liée à la distance considérée. Ces méthodes sont très simples et souvent utilisées, mais la limite majeure de cette approche est son manque de souplesse par rapport aux différents degrés de variance dans la réponse spectrale des données. Cependant, c'est le meilleur choix si les sites d'entrainement ont une faible superficie ou s'ils sont peu sûrs.

Basée sur la théorie de Bayes, la méthode de **classification par le Maximum de Vraisemblance** (**Maximum Likelihood** en anglais) est parmi les méthodes les plus utilisées en classification d'images satellitaires [Richards, 1993]. Dans cette méthode, il est question du calcul d'une fonction de probabilité multidimensionnelle, qui permet de déterminer la probabilité de chaque pixel d'appartenir à une classe donnée.
Le pixel est attribué à la catégorie pour laquelle cette probabilité est la plus grande.
La classification est effectuée à partir des matrices de covariances calculées pour chaque classe d'apprentissage entre les canaux utilisés.
En utilisant la formule de Bayes, on peut déduire la probabilité qu'un nouvel objet appartienne à une classe C_k, par l'intermédiaire des probabilités à postériori $P(C_k \mid x)$:

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)} \tag{2.1}$$

Malgré sa popularité, la classification par Maximum de Vraisemblance délivre quelques erreurs dans le cas où les échantillons d'apprentissage ne sont pas suffisants pour représenter les classes finales, la distribution des observations ne suit pas une loi de distribution gaussienne et enfin lorsque les différentes classes sont peu séparables.

2.4.2 Méthodes géométriques ou non paramétriques

Ces méthodes sont souvent sollicitées pour le faible coût en terme de temps de calcul. Ils ne tiennent pas compte de l'existence de la loi de probabilité suivie par les pixels de l'image. Parmi les méthodes géométriques les plus utilisées, la méthode des **k-plus proches voisins (k-ppv)** ou encore **k-NN** (**k Nearest Neighbours** en anglais) est l'une des premières méthodes non-paramétriques de classification développées [Cover et Hart, 1967]. C'est un moyen simple d'estimation non paramétrique de densité. Pour estimer la densité r_i de la classe C_i au point x, on recherche les K (K fixé à l'avance) plus proches voisins de x dans un ensemble de référence (l'ensemble d'apprentissage dont l'affectation des individus est connu à priori).

L'estimation de la densité est donnée par :

$$r_i = \frac{k_i(x)}{n_i V(x)} \tag{2.2}$$

où
K_i(x) est le nombre de points de C_i appartenant aux kppv de x,
n_i est le cardinal de la classe C_i,
V(x) est le volume de la plus petite boule contenant les kppv de x.

Cette méthode se simplifie en méthode de discrimination par voisinage en affectant à x la classe majoritaire parmi les kppv. Le résultat de la discrimination dépend de la valeur de K. C'est pourquoi il est intéressant de faire varier K afin d'obtenir les meilleurs résultats possibles.

Les principaux inconvénients de cette méthode sont le coût de stockage (les éléments de l'ensemble d'apprentissage doivent être stockés) ainsi que le coût élevé de la recherche des k-ppv [Cocquerez et Philipp, 1995].

2.5 Classification non supervisée

Lorsqu'on ne dispose pas de données relatives à la réalité terrain, on parle de classification non supervisée. Dans ce cas, l'utilisateur n'intervient qu'une fois la classification effectuée, pour interpréter le résultat sans connaissance à priori.

Le principe de la méthode consiste à répartir les éléments de l'image en fonction de leur degré de ressemblance, dans les classes choisies (appelé cluster) suivant un critère statistique.

Parmi les algorithmes de classification non supervisée les plus populaires en traitement d'images satellitaires, nous trouvons l'algorithme des k-moyennes, connu en anglais sous le nom de k-means ainsi que l'algorithme ISODATA.

La notation suivante sera adoptée dans cette section :
X : L'ensemble de données d'entrée avec la moyenne $\overline{X}$.
N : Le nombre d'objets dans l'ensemble de données.
n_i : Le nombre d'objets dans le cluster i.
K : Le nombre de clusters.
c_i : Le centre du cluster i.
c_{pi} : La p^{ieme} composante du vecteur c_i.

2.5.1 Différents types de clustering

Le processus de clustering peut être formalisé de trois façons diférentes : le clustering dur, le clustering flou et le clustering doux. Le tableau 2.1 présente un exemple des degrés d'appartenance des objets aux clusters pour un résultat dur, doux et flou .

Le clustering dur (hard-clustering) est le plus utilisé. Il consiste à étiqueter un objet dans une et une seule classe.

Le clustering flou (fuzzy-clustering) spécifie le degré d'appartenance d'un élément à un certain groupe .

Le clustering doux (soft-clustering) appelé aussi le clustering par recouvrement. Il propose une affectation dure de chaque objet à une ou plusieurs classes. Il existe trés peu d'approches de clustering doux.

Résultat dur.

	C1	C2	C3
x1	1	0	0
x2	0	1	0
x3	0	0	1
x4	0	0	1

Résultat doux.

	C1	C2	C3
x1	1	1	0
x2	0	1	1
x3	0	1	1
x4	0	0	1

Résultat flou.

	C1	C2	C3
x1	0.9	0.1	0
x2	0	0.8	0.2
x3	0	0.3	0.7
x4	0	0	1.0

TABLE 2.1 – *Exemple des degrés d'appartenance des objets aux clusters pour un résultat dur, doux et flou.*

2.5.2 Algorithmes de clustering basés sur la distance

La grande majorité des algorithmes de clustering sont basés sur la notion de distance entre les données comme critère de similarité (ou de dissimilarité). Dans ce contexte, les algorithmes de clustering tentent souvent d'optimiser une fonction objective qui favorise des clusters à la fois compacts et bien séparés. Pour ces algorithmes, le choix de la fonction de distance est essentiel.

Il existe diverses distances, les plus courantes étant la distance euclidienne, la distance de Manhattan (city block en anglais) et la distance de Mahalanobis. Les deux première sont deux cas particuliers de la distance de Minkowski avec (r = 2 , r= 1) respectivement, et qui est donnée par :

$$d(x_1, x_2) = (\sum_{i=1}^{n} |x_{1,2} - x_{2,1}|^r)^{\frac{1}{r}} \ \ avec \quad r \geqslant 1 \tag{2.3}$$

La distance de Mahalanobis tient compte de la distribution statistique des données dans l'espace, c'est ce qui la différencie des autres distances. Elle est définie comme suit :

$$d(x_1, x_2) = \sqrt{(x_1 - x_2)' \sum{}^{-1} (x_1 - x_2)} \tag{2.4}$$

Avec $\sum^{-1}$: inverse de la matrice de covariance de x_1 et x_2.

2.5.3 Algorithme de type "Nuées Dynamiques" ISODATA

ISODATA [Ball et Hall, 1965] est acronyme pour Iterative Self-Organizing Data Analysis Technics A (le A étant ajouté pour rendre le mot prononçable). La méthode des nuées dynamiques consiste à calculer la partition optimale de l'ensemble des individus en sous-ensembles, chaque sous-ensemble (classe) étant représenté par un "noyau".

Basé sur l'algorithme des k-moyennes, l'algorithme ISODATA permet une évolution du nombre de classes durant l'exécution. En effet, il autorise, au cours des itérations, la fusion entre deux classes proches, la séparation entre deux classes si leur dispersion (évaluée par l'écart-type) dépasse un certain seuil et enfin la suppression des classes dont la fréquence de nombre de pixels est petite.

2.5.4 Algorithme de k-means

Proposé par [MacQueen, 1967], l'algorithme k-moyennes, appelé algorithme des centres mobiles, est un des plus simples algorithmes de classification automatique d'objets et le plus fréquemment utilisés en télédétection du fait qu'il est très facile à mettre en oeuvre [Zheng et al., 2008][Koonsanit et al., 2012][Usman, 2013][Kumar et al., 2016].

L'algorithme k-means ne nécessite que deux paramètres, qui sont supposés être fixes, le nombre de clusters k, et le nombre d'itérations. L'algorithme k-means est une procédure itérative. La première étape consiste à initialiser aléatoirement les centres des clusters. La deuxième étape consiste à affecter chaque pixel de l'image au cluster le plus proche. La troisième étape consiste à mettre à jour les centres des clusters en calculant la moyenne des vecteurs caractéristiques des pixels appartenant à ce cluster. Les deuxième et troisième étapes sont répétées jusqu'à ce que le nombre d'itérations soit atteint ou que la fonction d'erreur ne change pas de manière significative. Cet algorithme vise à minimiser la fonction objective suivante (équation(2.5)).

$$j = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2 \tag{2.5}$$

Où $||x_i^{(j)} - c_j||^2$ est une mesure de distance choisie entre un point de données $x_i^{(j)}$ et le centre de cluster Cj ; qui est un indicateur de la distance des n points de données de leurs centres de cluster respectifs.

L'algorithme k-means est donné par :

Entrées :
K le nombre de clusters désiré, d une mesure de dissimilarité sur l'ensemble des objets à traiter X.

Sortie :
Une partition C = $\{C_1, C_2, ..., C_k\}$.

Etape 0 :

1- Initialisation par tirage aléatoire dans X, de k centres $x^*_{1,0}, ..., x^*_{k,0}$.

2- Constitution d'une partition initiale $C_0 = \{C_1, C_2, ..., C_k\}$ par allocation de chaque objet $x_i \in X$ au centre le plus proche :

$$C_l = \left\{ x_i \in X / d(x_i, x^*_{l,0}) = \min_{h=1,...,k} d(x_i, x^*_{h,0}) \right\}$$

3- Calcul des centres des k classes obtenues $x^*_{1,1}, ..., x^*_{k,1}$.

Etape t :

4- Constitution d'une nouvelle partition $C_t = \{C_1, C_2, ..., C_k\}$ par allocation de chaque objet $x_i \in X$ au centre le plus proche :

$$C_l = \left\{ x_i \in X / d(x_i, x^*_{l,t}) = \min_{h=1,...,k} d(x_i, x^*_{h,t}) \right\}$$

5- Calcul des centres des k classes obtenues $x^*_{1,t+1}, ..., x^*_{k,t+1}$.

6- Répéter **les étapes 4 et 5** tant que des changements s'opèrent d'un schéma C_t à un schéma C_{t+1} ou jusqu'à un nombre τ d'itérations.

7- Retourner la partition finale C_{finale}.

Plusieurs variantes de l'algorithme existent dans la littérature, c'est le cas par exemple de FCM (Fuzzy C-Means)[Bezdek et al., 1984], des k-médoïdes [Kaufman et Rousseeuw, 1987], de BKM (Bisecting k-Means) [Steinbach et al., 2000] et de KHM (K-Harmonic Means) [Zhang, 2000].

2.5.5 Algorithme de Fuzzy C-Means

Parmi les algorithmes de classification floue on peut citer le « C-moyennes floues», nommé plus couramment « Fuzzy C-means (FCM) ». FCM est une extension directe de l'algorithme classique des k-moyennes (k-means) ; où l'on a introduit les ensembles flous dans la définition des classes. Cet algorithme a été introduit par F.C. Dunn [Dunn, 1973]. Il fut généralisé par J.C.

Bezdek [Bezdek et al., 1984]. L'objectif est de construire une partition floue de l'image traitée. En d'autres termes, chaque pixel peut appartenir simultanément à deux ou plusieurs clusters et aura une "valeur d'appartenance" pour chaque cluster. Pour cela, il est basé sur la minimisation de la fonction objective suivante (équation(2.6)) :

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} U_{ij}^{m} ||x_i - c_j||^2 \tag{2.6}$$

où
L'exposant m module flou dans la partition ($1 \leq m < \infty$).
U_{ij} est le degré d'appartenance de xi au cluster j.
x_i est la $i^{ème}$ donnée mesurée de d - dimension.
c_j est c_j est le centre à d- dimensions du cluster.
| |*| | est la distance entre le pixel k et le centre.

L'algorithme est donné par :

1- Initialiser la matrice U= $[u_{ij}]$, U^0.
2- **A l'étape k** : calculer les vecteurs centres $C^{(k)} = [c_j]$ avec $U^{(k)}$.

$$C_j = \frac{\sum_{i=1}^{N} U_{ij}^{m} X_i}{\sum_{i=1}^{N} U_{ij}^{m}}$$

3- Mise à jour de $U^{(k)}$, $U^{(k+1)}$.

$$U_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\left\| X_i - C_j \right\|}{\left\| X_i - C_k \right\|} \right)^{\frac{2}{m-1}}}$$

4- **Si** $\left\| U^{k+1} - U^{k} \right\| < \epsilon$ alors **STOP** ; **sinon** retour à **l'étape 2**.

2.5.6 Algorithme K-Harmonic means

L'algorithme de clustering K-Harmonic means (KHM) est une version améliorée du K-Means qui a été proposé par Zhang en 1999 et 2000 [Zhang, 2000] et modifié par Hammerly et Elkan en 2002 [Hamerly et Elkan, 2002]. La méthode KHM est moins sensible à la procédure d'initialisation que la méthode K-means. L'insensibilité à l'initialisation est attribuée à une fonction de pondération dynamique, qui augmente l'importance des points de données qui sont éloignés de tout centre dans l'itération suivante [Zhang, 2000].

Les étapes de l'algorithme KHM sont données par :

Étape 1 : Choisir K centres initiaux c_j (j = 1...K) parmi N points de données et lancer KHM^* = 0.

Étape 2 : Calculer la valeur de la fonction de performance KHM (X) définie comme suit :

$$KHM(X) = \sum_{i=1}^{N} \left(\frac{k}{\sum_{j=1}^{k} \frac{1}{\left\| X_i - C_j \right\|^q}} \right)$$

Où : X_j désigne un objet dans l'ensemble de données d'entrée, q est un paramètre et laissez $q \geq 2$.

Étape 3 : Calculer T_{ij} (i= 1...N, j= 1...K) les éléments selon l'équation suivante :

$$T_{ij} = \frac{\left\| X_i - C_j \right\|^{-q-2}}{\sum_{j=1}^{K} \left\| (X_i - C_j) \right\|^{-q-2}}$$

Étape 4 : Obtenir le poids L_i de chaque point de données donné par :

$$L_i = \frac{\sum_{j=1}^{K} \left\| X_i - C_j \right\|^{-q-2}}{(\sum_{j=1}^{K} \left\| (X_i - C_j) \right\|^{-q})^2}$$

Étape 5 : Mettez à jour chaque centre de cluster comme suit

$$C_j = \frac{\sum_{i=1}^{N} T_{ij} L_i X_i}{\sum_{i=1}^{N} T_{ij} L_i}$$

Étape 6 : Si $|KHM^* - KHM| > \epsilon$, alors KHM^* = KHM et retournez à **l'étape 2** ; sinon passez à **l'étape 7**.

Étape 7 : Affecter chaque point de données X_i au cluster le plus proche C_j comme suit :

$$j = \arg \underset{j=1...k}{max} T_{ij}$$

2.5.7 Algorithme Bisecting K-means

Introduit par Steinbach en 2000 [Steinbach et al., 2000], l'algorithme BKM est une version améliorée de l'algorithme K-means classique qui offre un gain en termes de temps de calcul tout en préservant une bonne précision des résultats. L'idée du BKM est de diviser successivement par deux un ensemble de données jusqu'à ce que le nombre souhaité de clusters soit atteint. A chaque itération un groupe est sélectionné pour être divisé à son tour. Plusieurs manières dédiées au choix de ce dernier ont été utilisées. Par exemple, en choisissant le plus grand cluster en termes de cardinalité ou le cluster ayant la plus petite SSE (Somme des erreurs au carré).

$$SSE = \sum_{j=1}^{i} \sum_{x \in C_j} \left\| x - c_j \right\|^2 \tag{2.7}$$

$$c_j = \frac{1}{m_j} \sum_{x \in Cj} x \tag{2.8}$$

L'équation (2.7) sert de fonction objective de la procédure de clustering, où i désigne le nombre actuel de clusters, et x est un membre du $j^{ième}$ cluster C_j (j = 1, 2, ..., i). Pour l'équation (2.8), c_j est le centroïde de C_j qui contient m_j membres. Un SSE plus petit est corrélé avec un meilleur résultat de clustering. Par conséquent, l'objectif est de déterminer une méthode de clustering qui minimise l'SSE.

Le pseudo-code ci-dessous décrit la procédure de BKM.

Entrée : Un ensemble de données, un nombre de cluster K.

Initialiser une Table de Cluster (TC) contenant toutes les données d'un cluster ;

Pour chaque cluster **faire**
Initialiser SSE ;

  Pour chaque cluster existant C_j, **faire**
    Diviser C_j en 2 clusters par K-means ;
    Calculer SSE(j) pour les (i + 1) clusters actuels ;
  Fin
Choisir le cluster C_m avec un SSE minimum ;
Diviser C_m en 2 clusters par K-means et ajouter ces nouveaux clusters dans TC ;
Fin

Sortie :
Données avec K clusters.

2.6 Méthodes neuronales

Les réseaux de neurones sont à l'origine d'une tentative de modélisation mathématique du cerveau humain. Le principe général consiste à définir des unités simples appelées neurones. Chacune étant capable de réaliser quelques calculs élémentaires sur des données numériques. On relie ensuite un nombre important de ces unités formant ainsi un outil de calcul puissant. Plusieurs modèles ont été proposés et tentent de résoudre des problèmes insolubles avec des méthodes informatiques traditionnelles, ou au moins de rendre le coût de cette résolution acceptable selon certains critères. En télédétection spatiale, l'approche par réseaux de neurones intègre, en plus de données satellitaires, des données issues de sources différentes (altitude, pente, ..etc.) sans se soucier de leur nature ou de leur distribution. Les récents progrès en apprentissage automatique ont montré les très grandes performances des réseaux de neurones convolutifs pour de nombreuses applications, y compris la classification d'images aériennes et satellites [Yu et al., 2020][Thirupathi et Nagasudha, 2020][Karthik et Sangeetha, 2021].

2.6.1 Algorithme Self-Organizing Map

La carte auto-organisatrice (SOM) (Self-Organizing Map en anglais) introduite par Teuvo Kohonen en 1995 [Kohonen, 1995] est un réseau neuronal non supervisé. L'algorithme SOM est basé sur l'apprentissage compétitif. Le SOM offre une technique de visualisation des données qui aide à comprendre les données à haute dimension en réduisant la dimension des données en unités de carte. Le regroupement est effectué en mettant en concurrence plusieurs unités cartographiques ou neurones pour l'objet en cours. Le réseau de neurones artificiels est formé en fournissant des informations sur les entrées une fois que les données ont été saisies dans le système. L'unité gagnante ou active devient le vecteur de poids de l'unité la plus proche de l'objet actuel. Pour conserver les relations de voisinage dans l'ensemble des données d'entrée, les valeurs des variables d'entrée sont progressivement ajustées au cours de la phase d'apprentissage. Le SOM est également capable de généraliser. La capacité de généralisation garantit que le réseau peut comprendre ou classer des entrées qu'il n'a jamais rencontrées auparavant. Donc le processus de cartographie auto-organisatrice de Kohonen (Figure 2.4) est le suivant :

1. Initialisation.
2. Compétition.
3. Coopération.
4. Adaptation.

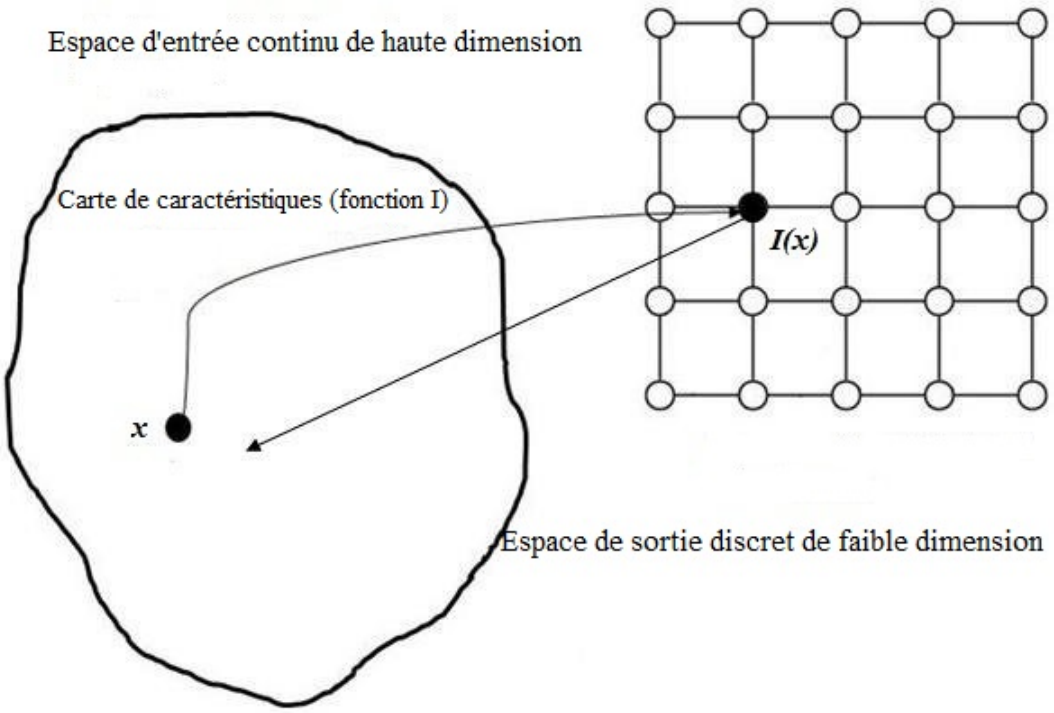


FIGURE 2.4 – *Topologie de l'algorithme SOM.*

Le SOM est constitué de m neurones situés sur une carte régulière de faible dimension, généralement une carte bidimensionnelle. Ces neurones sont connectés à leurs voisins selon des connexions topologiques. Chaque neurone i possède un vecteur de poids à d dimensions w = ($w_{i1}, w_{i2}, ..., w_{id}$), où i = 1, 2, ..., m, qui a la même dimension que l'espace d'entrée.

Les étapes de l'algorithme SOM sont données par :

(a) Initialiser les vecteurs de poids w_i des m x n neurones.

(b) Sélectionner aléatoirement un vecteur d'entrée x(t) et le transmettre à tous les neurones en même temps et en parallèle.

(c) Trouvez le neurone gagnant c, le neurone gagnant est appelé l'unité de meilleure correspondance BMU (Best Matching Unit), en utilisant l'équation suivante :

$$c = arg(\min_{1 \leqslant i \leq mn} \{\|w_i(t) - x(t)\|\})$$

$\|.\|$ est la mesure de la distance euclidienne.

Où $x(t)$ et $w_i(t)$ sont respectivement le vecteur d'entrée et le vecteur de poids du neurone i à l'itération t.

(d) Le vecteur de poids des neurones est mis à jour à l'aide de l'équation suivante :

$$w_i(t+1) = w_i(t) + h_{c,i}(t)\,[x(t) - w_i(t)]$$

Où $h_{c,i}(t)$ est une fonction de voisinage gaussienne donnée ci-dessous :

$$h_{c,i}(t) = \alpha(t).exp(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)})$$

Où r est la position des coordonnées du neurone sur la carte, $\alpha(t)$ est le taux d'apprentissage et $\sigma(t)$ est la largeur du rayon du voisinage.
$\alpha(t)$ et $\sigma(t)$ diminuent tous deux selon les équations suivantes :

$$\alpha(t) = \alpha(0)(\frac{\alpha(T)}{\alpha(0)})^{\frac{t}{T}}$$

$$\sigma(t) = \sigma(0)(\frac{\sigma(T)}{\sigma(0)})^{\frac{t}{T}}$$

Où T est la longueur d'apprentissage.

(e) Pour toutes les données d'entrée, les étapes **(b)** à **(d)** sont répétées.

2.7 Problèmes et limites du clustering

Malgré l'existance d'un grand nombre de méthodes de clustering ainsi que leur utilisation avec succès dans de nombreux domaines, le clustering pose encore de nombreux problèmes. Ces problèmes sont liées d'une part à l'augmentation de la quantité de données détectées à distance et leur hétérogénéité mais également au faite que chaque algorithme de clustering nécessite un certain nombre de paramètres et le plus important d'entre eux est le nombre de clusters, que l'utilisateur doit définir a priori. Un choix approprié de ce dernier peut générer des résultats de clustering pauvres.

Dans la section suivante, nous allons voir les principales mesures de qualité qui permettent d'évaluer un clustering.

2.8 Critères d'évaluation de la qualité d'un clustering

Les trois questions fondamentales qui doivent être abordées dans tout scénario typique de clustering sont les suivantes :
(i) quel est le bon clustering d'un ensemble de données en entrée.
(ii) combien de clusters sont réellement présents dans l'ensemble de données ?
(iii) et quel indice à appliquer pour avoir un bon clustering ?

L'évaluation de la qualité des résultats du clustering est une tâche difficile qui fait l'objet de recherches actives depuis des années [Jain et Dubes, 1988][Halkidi et al., 2001][Pakhira et al., 2004][Tan et al., 2005][Zhao, 2012][Arbelaitz et al., 2013], de nouvelles méthodes étant régulièrement proposées. La principale difficulté de l'évaluation des résultats de clustering réside dans la nature non supervisée inhérente au clustering lui-même et dans l'absence de consensus sur ce que devrait être un "bon clustering". Dans ce contexte, l'évaluation d'un résultat de clustering est toujours plus ou moins subjective, chaque critère d'évaluation favorisant un concept d'un bon clustering (forme, séparation compacte, etc.) par rapport aux autres. Par conséquent, les notions de bon clustering et de meilleur clustering dépendront à la fois du critère d'évaluation et de l'algorithme de clustering, certains critères d'évaluation favorisant certains algorithmes plutôt que d'autres.

Pourtant, malgré cette subjectivité relative, il existe un large éventail de critères d'évaluation couramment utilisés en apprentissage automatique pour évaluer et comparer les résultats de clustering. Il existe plusieurs taxonomies disponibles dans la littérature pour ces critères d'évaluation [Halkidi et al., 2001][Jain et Dubes, 1988][Tan et al., 2005], la plupart d'entre eux se divisant en 3 groupes distincts :

Les indices non supervisés, également appelés **indices internes** : ils utilisent uniquement les informations internes des données ainsi que les

caractéristiques des clusters.

Les indices supervisés, également appelés **indices externes** : ils évaluent le degré de similarité entre une solution de clustering et une partition connue du jeu de données (parfois appelée vérité terrain).

Les indices relatifs : ils constituent une classe distincte de critères qui permettent de comparer plusieurs résultats de clustering d'un même algorithme. Les indices relatifs utilisent simplement des critères externes et internes.

2.8.1 Indices de validité interne

Dans le cadre de la classification non supervisée, on définit une bonne partition comme étant la partition qui permet de réaliser un bon compromis entre l'inertie inter-clusters et l'inertie intra-clusters [Handl et al., 2005]. La plupart des indices de validité interne quantifient la qualité d'un partitionnement particulier en termes de compacité et de séparation entre les clusters :

- **L'inertie inter-clusters** : aussi appelée séparabilité, elle permet de quantifier la dispersion des clusters, en d'autres termes, elle mesure la dissimilarité moyenne entre les clusters, pondérée par leurs fréquences. Dans la plupart des problèmes de clustering, cette mesure est à maximiser.

- **L'inertie intra-cluster** : aussi appelée compacité (cohésion), elle permet de mesurer la distance moyenne entre les objets d'un même cluster. Dans ce cas-là, on cherche à minimiser l'inertie de manière à obtenir des clusters groupant des objets ayant des valeurs les plus similaires possibles.

Une catégorie d'indices internes est basée sur ces propriétés, nous présentons dans la suite les mesures d'évaluation les plus connues.

Erreur Quadratique Moyenne (EQM)

EQM est l'un des moyens les plus simples d'évaluer la qualité d'un résultat pour les algorithmes de clustering qui utilisent des centroïdes. Étant donné une solution de clustering S avec K clusters, elle peut être calculée comme indiqué dans l'équation (2.9) où d(*) est une fonction de distance, $|c_i|$ est le nombre d'éléments liés au cluster c_i et μ_k le centroïde d'un cluster c_k.

$$EQM = \frac{1}{\sum_{i=1}^{K} |C_i|} \sum_{k=1}^{K} \sum_{x \in C_k} d(x - \mu k)^2 \tag{2.9}$$

Pour qu'un résultat de clustering soit considéré comme bon, l'EQM doit être aussi faible que possible.

Indice de Dunn

L'indice de Dunn [Dunn, 1973] est un autre critère interne représenté par l'équation (2.10) où D(c_i, c_j) est une métrique de distance entre deux clusters c_i et c_j.
Δ_i est une mesure de dispersion pour un cluster c_i.

$$Dunn = \frac{min_{i \neq j} D(c_i, c_j)}{max_{i \in [1...K]} \Delta_i} \tag{2.10}$$

Un indice de Dunn plus élevé indique un meilleur clustering.

Indice de Davies-Bouldin (DB)

Cet indice tient compte à la fois de la compacité et de la séparabilité des groupes [Davies et Bouldin, 1979], il représenté par l'équation (2.11) . Une valeur faible de l'indice DB indique un clustering de bonne qualité.

$$DB = \frac{1}{K} \sum_{i=1}^{K} \underset{j \neq i}{max} \frac{\Delta_i + \Delta_j}{D(c_i, c_j)} \tag{2.11}$$

Bayesian information Criterion (BIC)

C'est un critère de sélection de modèles parmi un ensemble fini de modèles [Akaike, 1974], une valeur élévée de l'indice BIC indique un clustering de bonne qualité. Il est basé, en partie, sur la fonction de vraisemblance et il est formulé comme suit :

$$BIC = \sum_{i=1}^{M} (n_i log \frac{n_i}{N} - \frac{n_i \times D}{2} log(2\pi) - \frac{n_i}{2} log \sum_i - \frac{n_i - M}{2}) - \frac{1}{2} M log N \tag{2.12}$$

où

$$\sum_i = \frac{1}{N - M} \sum_{j=1}^{n_i} \left\| X_j - C_i \right\|^2 \tag{2.13}$$

Indice de Silhouette (IS)

L'indice de Silhouette [Rousseeuw et Leroy, 1987] est un autre critère interne qui évalue la compacité des clusters et leur séparation ou non. La principale différence entre l'indice de Silhouette et l'indice de Dunn ou l'indice de Davies-Bouldin est la suivante : l'indice de Silhouette peut être calculé pour un objet donné x, un cluster donné c_i, ou pour l'ensemble du clustering C.

Pour un élément de données, l'indice de silhouette est défini comme indiqué dans l'équation (2.14) où a_x est la distance moyenne entre l'objet observé x et tous les autres objets qui appartiennent au même cluster que x, et b_x est

la distance moyenne entre x et tous les autres objets qui ne sont pas dans le même cluster que x.

$$IS(x) = \frac{b_x - a_x}{max(a_x, b_x)} \tag{2.14}$$

L'indice Silhouette prend des valeurs comprises entre -1 et 1. Une valeur positive ($a_x < b_x$) signifie que x est plus proche des objets appartenant à ses clusters que des objets appartenant à d'autres clusters. Par conséquent, une valeur positive proche de 1 signifie que x est probablement dans le bon cluster, tandis qu'une valeur négative signifie que x devrait être dans un autre cluster.

Indice WSJI

Cet indice [Sun et al., 2004] utlise une combinaison linéaire de compacité et de séparation de la compacité et de la séparation floues pour mesurer le clustering. il est défini par l'équation suivante :

$$WSJI(K) = Scat(K) + \frac{Sep(K)}{Sep(K_{max})} \tag{2.15}$$

où

$$Scat(K) = \frac{\frac{1}{K}\sum_{i=1}^{K} \|\sigma(Z_i)\|}{\|\sigma(X)\|} \tag{2.16}$$

$$Sep(K) = \frac{D_{max}^2}{D_{min}^2} \sum_{i=1}^{K} (\sum_{k=1}^{K} \|Z_i - Z_k\|^2)^{-1} \tag{2.17}$$

$$D_{max} = max\|Z_i - Z_k\| \tag{2.18}$$

$$D_{min} = min\|Z_i - Z_k\| \tag{2.19}$$

Sep(K_{max}) fait référence au Sep(K) avec le nombre maximum de clusters. Une valeur minimale de l'indice WSJI signifie un bonne qualité de clustering.

Indice WB

L'indice WB est défini comme le rapport entre la mesure de la compacité du cluster (Sum-of-Squares **W**ithin (SSW)) et sa mesure de séparation (Sum-of-Squares **B**etween (SSB)) [Zhao, 2012][Zhao et Fränti, 2014]. Il est donné par l'équation (2.20) :

$$WB = K\frac{\sum_{i=1}^{N} \|x_i - c_{pi}\|^2}{\sum_{i=1}^{K} n_i \|c_i - \bar{X}\|} \tag{2.20}$$

Une valeur minimale de l'indice WB signifie une bonne qualité de clustering.
On s'intéresse, dans ce qui suit, à cet indice qui sera utilisé le long de ce travail.

2.8.2 Indices de validité externe

Lorsque les classes des objets réels sont connues, il est possible de comparer le résultat d'un clustering avec la partition réelle. Bien que ces critères externes ne soient pas des indices de clustering appropriés, c'est le moyen le plus pratique d'évaluer un nouvel algorithme de clustering en l'appliquant à un ensemble de données dont les classes ou les clusters sont connus. Les indices qui permettent d'évaluer un clustering sur la base de la comparaison entre une partition de clustering et les classes réelles sont appelés indices supervisés ou critères externes, car ils s'appuient sur des informations qui ne proviennent ni des données ni du clustering mais d'une vérité terrain externe utilisée à des fins de comparaison.

Différents critères ont été proposés dans la littérature [Davis et Goadrich, 2006][Fawcett, 2006].

Ces critères consistent à comparer deux partitions C_1 et C_2. La première chose à savoir lors de la comparaison de deux partitions est que dans la plupart des cas, il n'est pas possible d'établir un lien direct entre les clusters des deux partitions, ou dans notre cas entre les clusters et les classes réelles. En raison de ce problème, les indices externes ne sont pas directement basés sur les classes et les clusters des objets. Au lieu de cela, ils comparent des paires d'objets pour voir s'ils sont mis ensemble ou séparés dans les deux partitions.

Soit
aa le nombre de paires d'objets qui sont dans la même classe en C_1 et C_2,
bb le nombre de paires d'objets qui sont dans des classes différentes en C_1 et C_2,
ab le nombre de paires d'objets qui sont dans la même classe en C_1 mais pas en C_2,
et **ba** le nombre de paires d'objets qui sont dans des classes différentes en C_1 mais dans la même classe en C_2.

Plus **aa** et **bb** sont élevés, plus les deux partitions sont similaires. Nous introduisons maintenant plusieurs critères basés sur ces nombres de paires.

Précision

La précision évalue la probabilité que deux objets soient dans la même classe dans la partition C_2 sachant qu'ils sont dans la même classe dans la partition C_1. La précision prend ses valeurs entre 0 et 1.

$$Precision = \frac{aa}{aa + ab} \tag{2.21}$$

Rappel

Le rappel est la probabilité pour que deux objets soient dans la même classe dans C_1 s'ils le sont dans C_2. Le rappel prend ses valeurs entre 0 et 1.

$$Rappel = \frac{aa}{aa + ba} \tag{2.22}$$

Indice de Jaccard

L'indice de Jaccard prend ses valeurs entre 0 et 1 et est égal à 1 si et seulement si les deux partitions C_1 et C_2 sont identiques.

$$Jaccard = \frac{aa}{aa + ab + ba} \tag{2.23}$$

Indice de Rand

L'indice de Rand prend des valeurs entre 0 et 1 et est égal à 1 si les deux partitions C_1 et C_2 sont identiques.

$$Rand = \frac{aa + bb}{aa + ab + ba + bb} \tag{2.24}$$

F-Mesure

F- mesure, également connu sous le nom de F1 score, est un autre indice externe basé sur la précision et le rappel. F- mesure prend également ses valeurs dans [0, 1], 1 étant la meilleure valeur.

$$F - mesure = \frac{2 \times Precision \times Rappel}{Precision + Rappel} \tag{2.25}$$

Indice Kappa

L'indice Kappa, est défini par :

$$Kappa = \frac{P_0 - P_e}{1 - P_e} \tag{2.26}$$

avec

$P_0 = \frac{aa+bb}{aa+ab+ba+bb}$

$P_e = \frac{1}{(aa+ab+ba+bb)^2}(aa + ab) \times (aa + ba) + (ab + bb) \times (ba + bb)$

L'indice Kappa prend ses valeurs entre -1 et 1, 1 signifiant que les partitions C_1 et C_2 sont identiques.

Conclusion

Dans ce chapitre, nous avons présenté un aperçu sur les différentes approches utilisées pour la classification supervisées et non supervisées, les différents critères d'évaluations du clustering ainsi que les différents algorithmes de clustering appliqué à notre étude.

Cependant, il devient difficile d'obtenir des résultats de classification pertinents en utilisant un seul algorithme de regroupement. Pour tenter de résoudre ce problème nous allons voir dans le chapitre suivant comment plusieurs algorithmes de clustering peuvent être utilisés de manière simultanée pour produire des meilleurs résultats. Ainsi, nous proposons d'utiliser le paradigme des classificateurs combinés pour tirer parti des informations provenant de plusieurs algorithmes de cluterings différents.

Ensemble de clustering 3

Sommaire

Dans ce chapitre, nous présentons le concept d'ensemble de clustering, un sujet récent dans le domaine de l'apprentissage automatique. Nous commençons par aborder l'origine du concept ensemble clustering, ses motivations, ses objectifs et ses applications. Ensuite, nous présentons l'architecture de notre approche.

3.1 Introduction

Au cours des deux dernières décennies, l'ensemble de clustering a montré un grand potentiel puisqu'il vise à combiner plusieurs modèles de clustering pour produire un meilleur résultat que celui des algorithmes de clustering individuels en termes de qualité.

3.2 Problématique

Il existe un grand nombre d'algorithmes de clustering qui peuvent donner des résultats très différents avec les mêmes données, et choisir entre plusieurs résultats de clustering est souvent problématique. Ce problème ne peut être résolu qu'en demandant à un expert de choisir la méthode la plus adaptée et les paramètres qui fonctionneront le mieux pour un ensemble de données spécifiques. Il s'agit d'une tâche très difficile qui peut avoir une grande influence sur les résultats. Prendre ce type de décision nécessite une connaissance approfondie des données à analyser, mais aussi du grand nombre d'algorithmes disponibles. De plus, même avec un bon expert ayant une bonne connaissance des données et des algorithmes, il est toujours difficile de faire les bons choix lorsqu'il s'agit de clustering.

Pour tenter de résoudre ce problème, la communauté scientifique s'est intéressée depuis plusieurs années à la combinaison de plusieurs méthodes de clustering. Son but est de produire un unique clustering à partir d'un ensemble de plusieurs clusterings.

Le marquis de Condorcet a été l'un des premiers à formaliser cette notion [Condorcet, 1785]. Dans son "théorème du jury de Condorcet", il formule la probabilité qu'un groupe d'individus parvienne à une solution exacte. L'hypothèse de la version la plus simple du théorème est qu'un groupe souhaite prendre une décision par un vote majoritaire. Tous les membres du groupe peuvent choisir entre différents choix - dont l'un est correct - et chaque votant a une probabilité indépendante p de voter pour la bonne décision. Le théorème stipule alors combien de votants doivent être inclus dans le groupe. La réponse dépend du fait que p est supérieur ou inférieur à $\frac{1}{2}$: si p est supérieur à $\frac{1}{2}$ (chaque électeur a plus de chances de voter correctement), alors l'ajout d'électeurs augmente la probabilité que la décision majoritaire soit correcte. À la limite, la probabilité que la majorité vote correctement s'approche de 1 lorsque le nombre de votants augmente. En revanche, si p est inférieur à $\frac{1}{2}$ (chaque électeur est plus susceptible de voter de manière incorrecte), l'ajout d'autres électeurs aggrave la situation : le jury optimal est composé d'un seul électeur.

Dans le tableau 3.1, nous montrons un exemple de ce type de vote où "□" est une bonne réponse et "■" une mauvaise réponse.

Classifieur	Prédiction	Précision
Classifieur 1	□■□■■□□□□■□□□□■	$\frac{10}{15} = 0.667$
Classifieur 2	■■□□□□■□□□□□■□■	$\frac{10}{15} = 0.667$
Classifieur 3	□□■□□■■□□□□□■■□	$\frac{10}{15} = 0.667$
Vote	□■□□□□■□□□□□■□■	$\frac{11}{15} = 0.733$

TABLE 3.1 – *Exemple d'un vote majoritaire dans l'apprentissage non supervisé [Kuncheva, 2008].*

Les approches basées sur cette idée ont été largement étudiées dans l'apprentissage supervisé [Schapire, 1990][Wolpert, 1992][Kittler et al., 1998] où elles ont donné naissance à l'apprentissage d'ensemble ou classification par la méthode d'ensembliste. Cependant, Concevoir une nouvelle approche de clustering d'ensemble n'est pas si simple. En fait, plusieurs raisons rendent cette tâche difficile. L'une d'elles est que le custering est une méthode non supervisée dans laquelle l'algorithme n'a aucune connaissance préalable de la classe de données. De plus, il est impossible d'effectuer des techniques de validation croisée pour ajuster les paramètres de l'algorithme de clustering (nombre de clusters, ...), il n'y a donc pas de directives pour les programmeurs pour choisir l'algorithme et sélectionner les paramètres appropriés pour un ensemble de données de paramètres. En raison de ces difficultés inhérentes que l'idée de combiner différents algorithmes de clustering est devenue populaire dans le domaine de l'apprentissage non supervisé.

3.3 PROCESSUS DE LA MÉTHODE D'ENSEMBLE DE CLUSTERING

Le cœur d'un ensemble de clustering est la fonction de consensus qui doit résoudre trois problèmes : comment combiner les différentes solutions de clustering ? comment surmonter le problème de correspondance des étiquettes ? et comment assurer un consensus symétrique par rapport à toutes les partitions d'entrée ?

De nombreuses approches ont été développées pour résoudre les problèmes de clustering par consensus. [Vega-Pons et Ruiz-Shulcloper, 2011] ont résumé le processus d'ensemble de clustering en deux étapes principales : **la génération** et **le consensus**. La figure3.1 illustre ce processus, dans lequel l'entrée est l'ensemble de données original et la sortie est le consensus de clustering.

Étape de génération : Il s'agit de la première étape du processus d'ensemble de clustering, où un certain nombre de membres de l'ensemble sont générés en utilisant des techniques de génération particulières. [Vega-Pons et Ruiz-Shulcloper, 2011] ont souligné qu'une plus grande variance dans l'ensemble des membres de l'ensemble signifie que plus d'informations sont disponibles pour la fonction de consensus. De plus, il n'y a pas de contraintes

sur la façon dont les membres de l'ensemble doivent être obtenus [Vega-Pons et Ruiz-Shulcloper, 2011]. Par conséquent, différentes stratégies peuvent être appliquées. Dans la littérature, plusieurs techniques de génération ont été utilisées pour générer des membres pour construire un ensemble ; plus de détails sur ces techniques peuvent être trouvés dans la section 3.4.

Étape de consensus : La deuxième étape est celle où les membres générés sont combinés en utilisant une fonction de consensus pour obtenir le résultat final du clustering. Le succès d'un ensemble de clustering repose sur le choix d'une fonction de consensus qui peut améliorer la qualité de la solution finale de clustering [Greene et Cunningham, 2006]. Par conséquent, un certain nombre de fonctions de consensus ont été proposées dans la littérature ; la section 3.5 passe en revue certaines fonctions de consensus courantes.

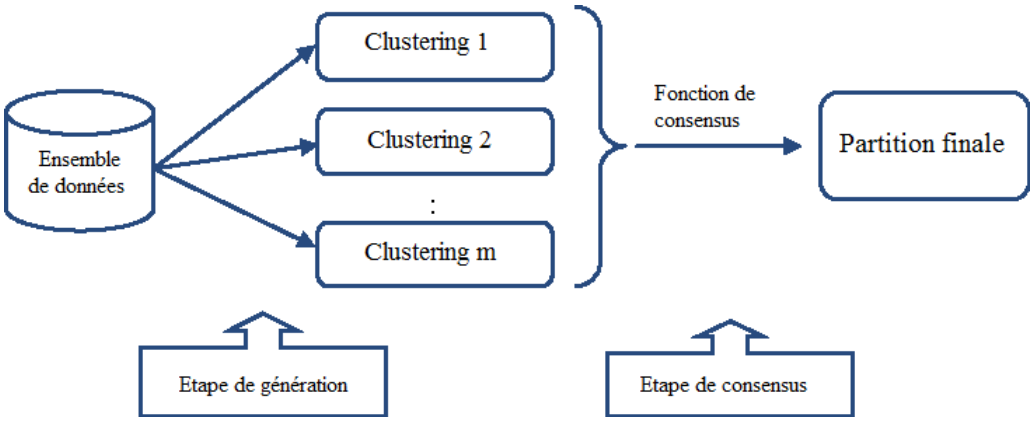


FIGURE 3.1 – *Processus d'ensemble de clustering [Vega-Pons et Ruiz-Shulcloper, 2011].*

3.4 TECHNIQUES DE GÉNÉRATION D'ENSEMBLES

[Iam-On et al., 2012] classent les techniques utilisées dans l'étape de génération en cinq catégories, comme le montre la figure 3.2 :

- **Ensemble homogène** : Un seul algorithme de clustering est exécuté pour générer un certain nombre de partitions avec plusieurs ensembles d'initialisations de paramètres.
- **Ensemble différent** : Chaque partition est générée avec un ensemble de cluster différent choisi aléatoirement.
- **Sous-espace de données ou sous-échantillon de données** : Chaque partition est générée par un sous-échantillon aléatoire des données, ou sur différents sous-espaces, ou en utilisant un sous-ensemble aléatoire de caractéristiques.
- **Ensemble hétérogène** : Chaque partition est générée en utilisant un algorithme de clustering différent.

– **Heuristique mixte** : Toute combinaison des techniques ci-dessus peut être utilisée pour générer un certain nombre de partition.

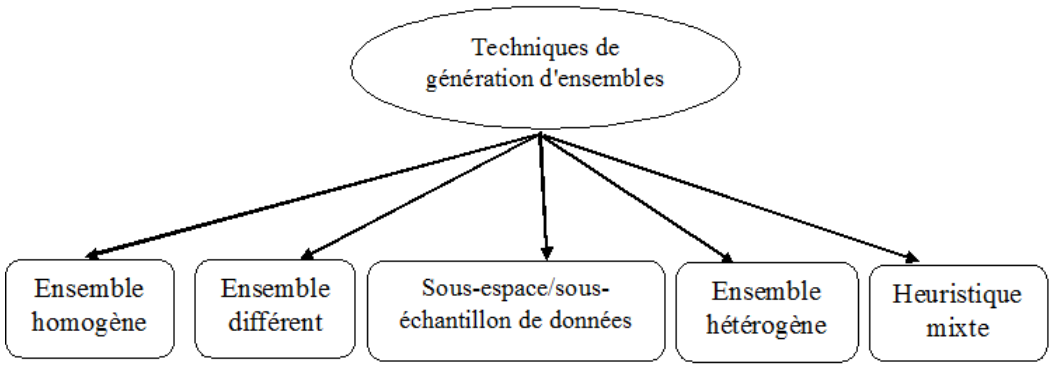


FIGURE 3.2 – *Approches de génération de partitions [Iam-On et al., 2012].*

3.5 FONCTIONS DE CONSENSUS

Des revues sur les méthodes d'ensemble de clustering peuvent être trouvées dans [Ghaemi et al., 2009][Ghosh et Acharya, 2011][Vega-Pons et Ruiz-Shulcloper, 2011][Yang, 2011] [Kuncheva, 2014][Sarumathi et al., 2014], où les auteurs ont essayé de classer ces méthodes selon leurs techniques. Parmi celles-ci, nous considérons le schéma de classification initialement proposé par [Vega-Pons et Ruiz-Shulcloper, 2011] en raison de sa simplicité. Cela facilite l'introduction des principales méthodes d'ensemble présentées dans la littérature. Ainsi, selon eux, la fonction de consensus peut être classée en deux approches principales : les approches basées sur la partition médiane et la cooccurrence d'objets.

3.5.1 Approche de la partition médiane

Cette approche traite la fonction de consensus comme un problème d'optimisation consistant à trouver la partition médiane par rapport à l'ensemble de clusters. La partition médiane est définie comme "la partition qui maximise la similarité avec toutes les partitions de l'ensemble de clustering" (Figure 3.3). Donc l'idée de base est de trouver une partition P qui maximise la similarité entre P et toutes les N partitions de l'ensemble : P_1, P_2, ..., P_N en utlisant la formule suivante (équation (3.1)) :

$$P = \arg \max_{P \in \mathbb{P}} \sum_{i=1}^{N} S(P, P_i) \tag{3.1}$$

où
P est la partition finale de l'agrégation.
$\mathbb{P}$ est l'ensemble des partitions possibles.

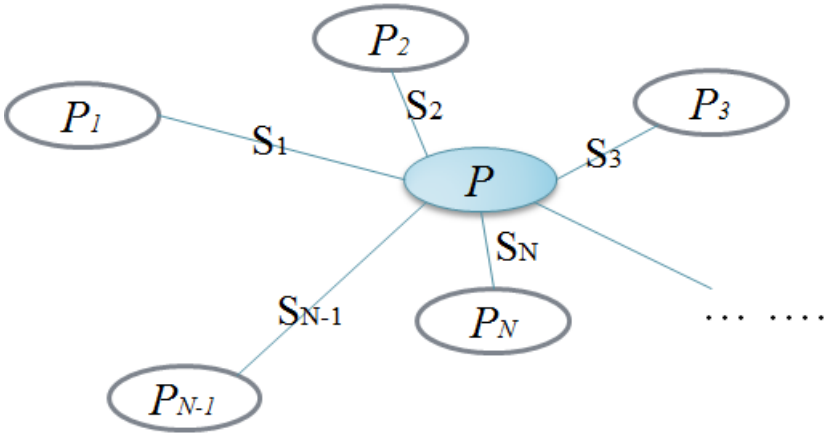


FIGURE 3.3 – *Principe de l'approche partition médiane.*

S(P_i, P_j) est une fonction de similarité entre deux positions arguments.

Parmi les fonctions de similarité entre partitions, on cite "Normalized mutual information" [Strehl et Ghosh, 2002], "Utility function" [Topchy et al., 2005], "Fowlkes-Mallows index" [Fowlkes et Mallows, 1983] et "Purity and inverse purity" [Zhao et Karypis, 2005].

Parmi les exemples de cette approche, on trouve la méthode basée sur la factorisation matricielle non négative [Li et al., 2007], la méthode génétique [*Yoon*^{a} et al., 2006][Luo et al., 2006] et la méthode basée sur les noyaux [Vega-Pons et al., 2010] .

Méthodes basées sur les algorithmes génétiques

Ces méthodes utilisent la capacité de recherche des algorithmes génétiques pour obtenir le clustering consensuel. En général, la population initiale est créée avec les partitions de l'ensemble de clusters et une fonction de fitness est appliquée pour déterminer quels chromosomes (partitions de l'ensemble d'objets) sont plus proches du clustering. Ensuite, des étapes de croisement et de mutation sont appliquées pour obtenir de nouveaux descendants et régénérer la population. Au cours de ce processus, si un critère de fin est atteint, la partition ayant la valeur de fitness la plus élevée est sélectionnée comme partition de consensus. Parmi les méthodes basées sur les algorithmes génétiques, on peut citer l'ensemble de regroupement hétérogène [*Yoon*^{a} et al., 2006] [*Yoon*^{b} et al., 2006]. La population initiale de cette méthode est obtenue en utilisant n'importe quel type de mécanisme de génération. Avec chaque paire de partitions obtenues à partir des objets, une paire ordonnée est créée. Le processus de reproduction utilise une fonction de fitness comme moyen unique de déterminer si une paire de partitions (chromosomes) survivra ou non à l'étape suivante. Dans cet algorithme, la

fonction d'aptitude est produite pour la comparaison de la quantité de chevauchements entre les partitions de chaque chromosome.

Méthodes basées sur la factorisation de matrices non négatives

[Li et al., 2007] a introduit une méthode d'ensemble de clustering basée sur un processus de factorisation de matrice non négative. La factorisation de matrices non négatives (Nonnegative Matrix Factorization (NMF)) [Cichocki et al., 2008] fait référence au problème de la factorisation d'une matrice de données M non négative donnée en deux facteurs matriciels, c'est-à-dire M≈AB, tout en exigeant que A et B soient non négatifs.

Dans cette méthode, on utilise tout d'abord la distance suivante entre les partitions (équation(3.2)) :

$$\mu(P, P^{'}) = \sum_{i,j=1}^{n} \mu_{ij}(P, P^{'}) \tag{3.2}$$

où $\mu_{ij}(P, P^{'}) = 1$ si x_i et x_j appartiennent au même cluster dans une partition et appartiennent à des clusters différents dans l'autre, sinon $\mu_{ij}(P, P^{'}) = 0$.

Cette méthode définit le regroupement consensuel comme le problème de la partition médiane, en fixant la distance (équation(3.2)) comme une mesure de ressemblance entre les partitions. La définition originale du problème est consécutivement relaxée pour transformer le problème en un problème d'optimisation qui peut être résolu par un processus itératif. Cependant, ce processus ne peut trouver que des minima locaux, plutôt qu'un minimum global du problème. Bien que les règles multiplicatives soient les techniques les plus courantes pour la factorisation matricielle non négative, il existe d'autres approches telles que les algorithmes des moindres carrés alternatifs à point fixe et les algorithmes Quasi-Newton qui peuvent être plus efficaces et obtenir de meilleurs résultats que les techniques multiplicatives [Vega-Pons et Ruiz-Shulcloper, 2011].

Méthodes basées sur les noyaux

[Vega-Pons et al., 2010] ont proposé l'algorithme Weighted Partition Consensus via Kernels (WPCK). Cet algorithme incorpore une étape intermédiaire, appelée analyse de la pertinence des partitions, dans la méthodologie traditionnelle des algorithmes d'ensembles de clustering, dans le but d'estimer l'importance de chaque partition avant le processus de combinaison. Dans cette étape intermédiaire, à chaque partition P_i est assignée une valeur de poids w_i qui représente la pertinence de la partition dans l'ensemble de clustering.

Dans cette méthode, la mesure de similarité suivante entre les partitions est définie comme suit :

$\widetilde{k} : \mathbb{P}_X \times \mathbb{P}_X \to [0,1]$ tel que :

$$\widetilde{k}(P_i, P_j) = \frac{k(P_i, P_j)}{\sqrt{k(P_i, P_i)k(P_j, P_j)}} \tag{3.3}$$

où la fonction k : $\mathbb{P}_X \times \mathbb{P}_X \to \mathbb{R}_+$ est donné par :

$$k(P_i, P_j) = \sum_{S \subseteq X} \delta_S^{P_i} \delta_S^{P_j} \mu(S \setminus P_i)\mu(S \setminus P_j) \tag{3.4}$$

avec

$$\delta_S^P = \begin{Bmatrix} 1 \, si \, \exists \, C \in P, \, S \subseteq C \\ 0 \qquad\qquad\quad sinon \end{Bmatrix} \tag{3.5}$$

et $\mu(S \setminus P)$ représente la signification du sous-ensemble S dans la partition P, qui peut être être calculée comme $\frac{|S|}{|C|}$ si $\exists C \in P$ tel que $S \subseteq C$.

3.5.2 Approche co-occurrence des objets

L'approche calcule d'abord la co-occurrence des objets dans les partitions, puis détermine leurs étiquettes de cluster pour produire un résultat consensuel. Fondamentalement, elle compte l'occurrence d'un objet dans un cluster, ou l'occurrence d'une paire d'objets dans le même cluster, et génère le résultat final du clustering par un processus de vote entre les objets. Parmi les exemples de ce type d'approche, on peut citer : la méthode de ré-étiquetage et de vote [Dudoit et Fridlyand, 2003][Zhou et Tang, 2006][Ayad et Kamel, 2010][Khedairia et Khadir, 2019], la matrice de co-association [Fred et Jain, 2005] et la méthode basée sur le graphique [Strehl et Ghosh, 2002][Fern et Brodley, 2004].

Méthode de ré-étiquetage et de vote

On l'appelle aussi l'approche directe ou relabeling. Pour les autres approches, il n'est pas nécessaire de résoudre explicitement le problème de correspondance entre les étiquettes des clusters connus et les dérivés. L'approche par le vote tente de résoudre le problème de correspondance, puis un simple vote peut être appliqué pour affecter les objets dans les clusters afin de déterminer la partition consensuelle finale. Cependant, la correspondance des étiquettes est exactement ce qui rend difficile la combinaison non supervisée. L'idée principale est de permuter les étiquettes des clusters de manière à obtenir le meilleur accord entre les étiquettes de deux partitions. Toutes les partitions de l'ensemble doivent être ré-étiquetées en fonction d'une partition de référence fixe. La partition de référence peut provenir de l'ensemble ou d'une nouvelle classification de l'ensemble de données. En outre, une procédure de vote significative suppose que le nombre de clusters dans chaque partition donnée est le même que dans la partition de référence. Cela nécessite que le nombre de clusters dans la partition de consensus de référence soit connu.

[Ayad et Kamel, 2010] ont une formulation générale du problème du vote en tant que problème de régression multiréponse. Parmi les méthodes basées sur le réétiquetage, on trouve le vote à la pluralité (Plurality Voting (PV)) [Fischer et Buhmann, 2003], le vote-fusion (Voting-Merging (V-M))[Dimitriadou et al., 2001], le vote pour le clustering flous [Dimitriadou et al., 2002], le vote des clusters actives (voting Active Clusters (VAC))[Tumer et Agogino, 2008], le vote cumulatif (Cumulative Voting (CV)) [Ayad et Kamel, 2008] et les méthodes proposées par [Zhou et Tang, 2006] et par [Gordon et Vichi, 2001].

Les travaux de [Fischer et Buhmann, 2003][Dudoit et Fridlyand, 2003][Hong et al., 2008], ont mis en œuvre une combinaison de partitions par le réétiquetage et par le vote. Leurs travaux ont suivi des approches directes de ré-étiquetage du problème de correspondance. Un relabeling peut être effectué de manière optimale entre deux clusters en utilisant l'algorithme de Hongrois (Hungarian algorithm). Après un ré-étiquetage globalement cohérent, le vote peut être appliqué pour déterminer l'appartenance de chaque objet à un groupe.

Notre approche est basée sur le principe de réétiquetage qui sera décrit dans la section 3.7.2.

Méthode de la matrice de co-association

La méthode la plus populaire basée sur la similarité par paire est la méthode de co-association. Son principe [Fred et Jain, 2005] est basé sur l'utilisation d'une matrice dite de co-association, cette matrice est calculée à partir des schémas de clustering en entrée (Figure 3.4), formellement exprimée comme suit :

$$co(x_i, x_j) = \frac{1}{M} \sum_{m=1}^{M} \delta(P_m(x_i), P_m(x_j)) \tag{3.6}$$

où x_i et x_j sont des objet, $P_m(x_i)$ représente l'étiquette associée de l'objet x_i dans la partition P_m et $\delta(a, b)$ est 1, si a = b, et 0 sinon. Il s'agit d'une matrice carrée dont la valeur de chaque position (i, j) de cette matrice est une mesure du nombre de fois où les objets x_i et x_j se trouvent dans le même cluster pour toutes les partitions de $\mathbb{P}$. Cette matrice peut être considérée comme une nouvelle mesure de similarité entre l'ensemble des objets X. Plus les objets x_i et x_j apparaissent dans les mêmes clusters, plus ils sont similaires. En utilisant la matrice de co-association "co" comme mesure de similarité entre les objets, la partition de consensus est obtenue en appliquant un algorithme de clustering.

Dans [Fred, 2001], un seuil fixe égal à 0,5 est utilisé pour générer la partition finale de consensus. Elle est obtenue en rejoignant dans le même cluster les objets dont la valeur de co-association est supérieure à 0,5.

Notez que la complexité de calcul des algorithmes de consensus basés sur la co-association est très élevée $O(n^2)$ et ne peut pas être appliqué à de grands ensembles de données [Topchy et al., 2003].

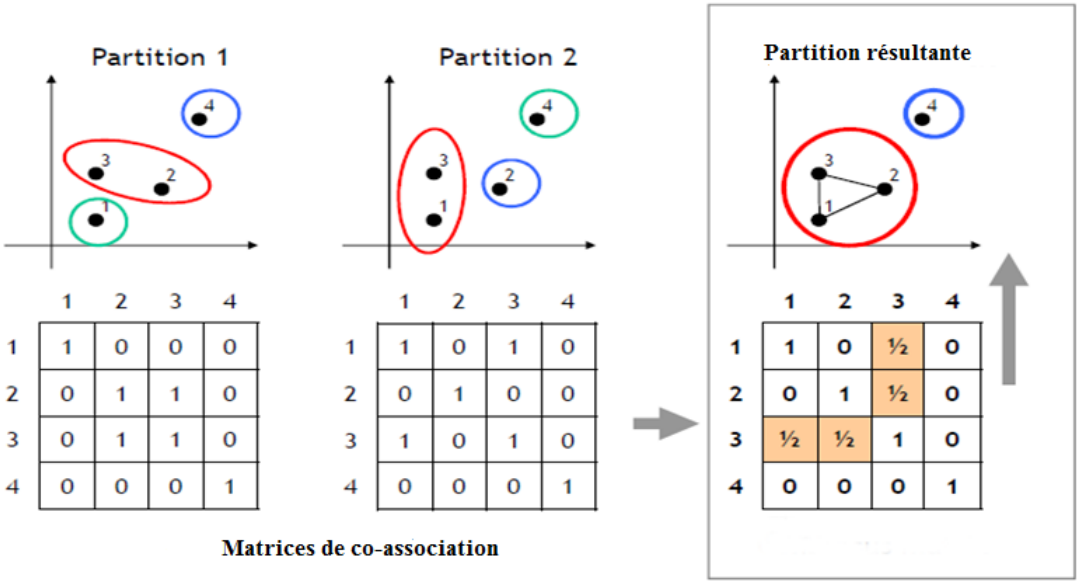


FIGURE 3.4 – *Principe de l'approche basée sur la matrice de co-association.*

Méthode basée sur le graphique

Les clusters pouvent être représentés comme des hyper-arêtes sur un graphe dont les sommets correspondent aux objets à regrouper, de sorte que chaque hyper-arête décrit un ensemble d'objets appartenant aux mêmes clusters (Figure 3.5). Ce type de méthodes d'ensemble de clustering transforme le problème de combinaison en un problème de partitionnement de graphes ou d'hypergraphes. Le problème du clustering consensuel se réduit alors à trouver la coupe minimale d'un hypergraphe. La coupe minimale de cet hypergraphe en k composantes donne la partition consensuelle requise [Strehl et Ghosh, 2002] [*Topchy*[b] et al., 2004].

Donc c'est construire un graphe pour représenter plusieurs résultats de clustering d'ensemble, puis trouver la partition optimale des données en minimisant la réduction du graphe [Fern et Brodley, 2004][Strehl et Ghosh, 2002]. La recherche de consensus est définie comme le résultat de recherche $P^{(*)}$ qui partage le plus d'informations avec les résultats de l'ensemble $\mathbb{P}$. Afin de trouver ce consensus, trois algorithmes (CSPA, HPGA, MCLA) proposés par [Strehl et Ghosh, 2002] utilisent le concept d'hypergraphe pour représenter l'ensemble des résultats de clustering.

Un hypergraphe est un ensemble de sommets et d'hyper-arêtes, une hyper-arête étant une généralisation du concept d'arête pouvant être connectée à un ensemble de sommets. Pour chaque résultat $P^{(i)} \in \mathbb{P}$, une matrice binaire d'appartenance $H^{(i)}$ est créée, composée d'une colonne pour chaque cluster du résultat (Tableau 3.2). La concaténation de l'ensemble des matrices H = $(H^{(1)}...H^{(N)})$ représente la matrice d'adjacence d'un hypergraphe à N sommets $\sum_{i=1}^{N} K^{(i)}$ hyper-arêtes. Chaque colonne h_i définit une hyper-arête où 1 indique que cette hyper-arête contient le sommet et 0 qu'elle ne le contient pas. Les trois algorithmes proposés utilisent cette représentation.

Partitions de clustering			
	$P^{(1)}$	$P^{(2)}$	$P^{(3)}$
x_1	1	1	1
x_2	1	2	2
x_3	2	1	1
x_4	2	2	2
x_5	3	3	3
x_6	3	4	3

Matrice d'appartenance										
	$H^{(1)}$			$H^{(2)}$				$H^{(3)}$		
	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}
x_1	1	0	0	1	0	0	0	1	0	0
x_2	1	0	0	0	1	0	0	0	1	0
x_3	0	1	0	1	0	0	0	1	0	0
x_4	0	1	0	0	1	0	0	0	1	0
x_5	0	0	1	0	0	1	0	0	0	1
x_6	0	0	1	0	0	0	1	0	0	1

TABLE 3.2 – *Exemple de représentation par hypergraphe.*

Dans le premier algorithme CSPA (Cluster-based Similarity Partitioning Algorithm), une matrice de similarité n x n (n étant le nombre d'objets) est construite à partir de l'hypergraphe. Celle-ci peut être considérée comme la matrice d'adjacence d'un graphe entièrement connecté, où les nœuds sont les éléments de l'ensemble X et où une arête entre deux objets a un poids associé égal au nombre de fois où les objets sont dans le même cluster. Ensuite, algorithme de partitionnement du graphe est utilisé pour obtenir la partition de consensus.

Dans le second algorithme à savoir HPGA (HyperGraph Partitioning Algorithm) partitionne directement l'hypergraphe, en créant un clustering consensus qui coupe le moins d'hyper-arêtes.

Et pour le dernière approche MCLA (Meta-CLustering Algorithm) définit d'abord la similarité entre deux clusters en termes de quantité d'objets groupés dans les deux, en utilisant l'indice de Jaccard. Ensuite, une matrice de similarité entre les clusters est construit, qui représente la matrice d'adjacence du graphe appelé méta-groupes.

Dans une autre méthode proposée par [Fern et Brodley, 2004], appelée HBGF (Hybrid Bipartite Graph Formulation), le problème se réduit à trouver une partition d'un graphe biparti pour former un consensus. L'algorithme HBGF combine les deux algorithmes CSPA et MCLA et exprime l'ensemble à travers un graphe bipartite où les instances et les clusters forment des sommets.

	P_1	P_2	P_3
v_1	1	1	1
v_2	1	2	2
v_3	2	1	1
v_4	2	2	2
v_5	3	3	3
v_6	3	4	3

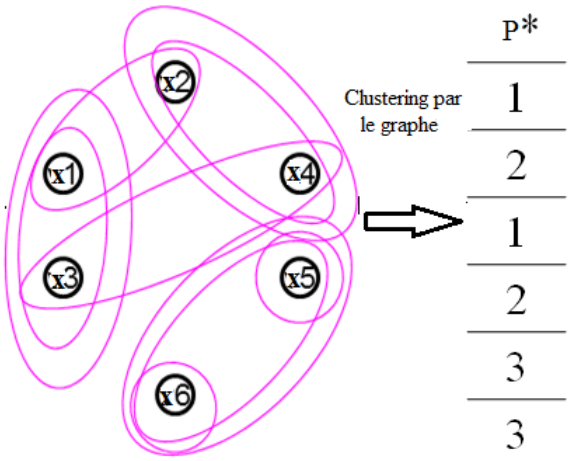


FIGURE 3.5 – *Exemple d'ensemble de clustering.*

3.6 APPLICATIONS D'ENSEMBLE DE CLUSTERING

- Réutilisation de connaissances : par exemple, utilisation d'une classification antérieure.
- Classification multivue : par exemple, classification d'images ou de documents web basée sur des annotations fournies par des utilisateurs.
- Fouille de données distribuée : par exemple, entreprises ayant des données commerciales privées.
- Amélioration de la qualité des résultats : par exemple, classification de données ayant des formes spécifiques.
- Production de solutions robustes : plusieurs exécutions d'un même algorithme initialisé différemment peuvent être utilisées, ce qui permet d'obtenir une solution plus robuste.

3.7 Approche proposée

Ensemble de clustering vise à générer plusieurs modèles et à les fusionner pour produire un cluster consensuel final [Chaouche Ramdane et al., 2019].

La figure 3.6 illustre la méthodologie adoptée de notre système à classificateurs multiple (MCS) [Chaouche Ramdane et al., 2022]. Elle se compose de deux étapes principales : la génération de partitions et la combinaison de partitions.

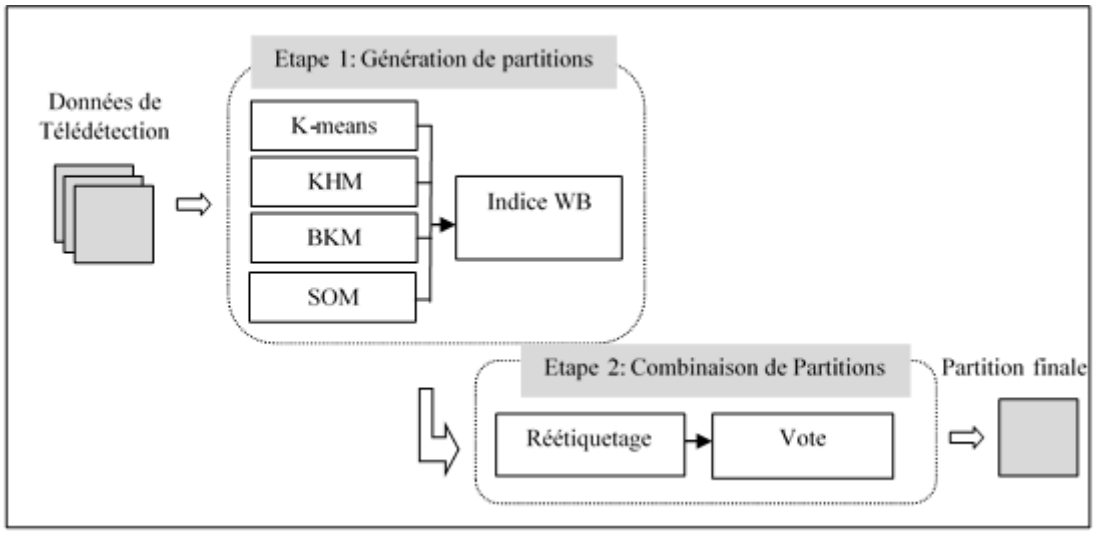


Figure 3.6 – *Architecture de l'approche proposée.*

3.7.1 Génération de partitions

Quatre algorithmes de clustering hétérogènes sont utilisés pour concevoir le MCS, à savoir l'algorithme bien connu K-means, l'algorithme K-harmonique (KHM), l'algorithme Bisecting K-means (BKM) et l'algorithme Self Organizing Map (SOM). Une description détaillée de ces algorithmes est donnée dans le chapitre 2.

Dans ce travail, nous avons proposé l'indice de validité interne WB pour estimer le nombre de clusters. En comparaison avec d'autres indices (indice DB, indice BIC, indice IS, indice WSJI, ...), une faible valeur de cet indice implique une simplicité et une haute qualité de clustering [Zhao, 2012][Zhao et Fränti, 2014]. Le choix de cet indice est basé sur plusieurs travaux réalisés en imagerie satellitaire [Mahi et al., 2015] [Labed et al., 2018], les résultats obtenus montrent l'efficacité de l'indice WB pour extraire le nombre optimal de clusters par rapport aux autres indices.

Pour déterminer le nombre optimal de clusters K, nous répétons successivement chaque algorithme de clustering pour un intervalle $[K_{min}, K_{max}]$ avec $K_{min} = 2$ et K_{max} est donné par la règle empirique $\sqrt{\frac{N}{2}}$ [Kodinariya et Makwana, 2013] et nous calculons la valeur de l'indice WB pour chacun des

résultats de partition correspondant à K. Ensuite, nous sélectionnons le K pour lequel la valeur WB est minimale.

3.7.2 Combinaison de partitions

Sachant que chaque algorithme de clustering peut donner un nombre différent de clusters en utilisant l'indice WB, il est nécessaire d'utiliser en plus un algorithme de vote pour trouver la correspondance entre les clusters d'étiquettes obtenus par les différentes partitions. L'approche de vote tente de résoudre le problème de correspondance. Ensuite, un schéma de vote simple tel que le vote majoritaire peut être utilisé pour assigner les objets dans les clusters afin de déterminer la partition consensuelle finale. Cependant, la correspondance des étiquettes est exactement ce qui rend difficile la combinaison non supervisée. L'idée principale est de permuter les étiquettes des clusters de manière à obtenir le meilleur accord entre les étiquettes de deux partitions. Toutes les partitions de l'ensemble doivent être ré-étiquetées en fonction d'une partition de référence fixe. La partition de référence peut être prise dans l'ensemble ou dans une nouvelle partition de l'ensemble de données [Yang, 2017].

Dans cette approche, le meilleur clustering qui fait office de référence est obtenu selon l'indice WB, et le nombre de clusters dans chaque partition d'entrée doit être le même que dans la partition de référence, la procédure entière implique simplement deux étapes :

Etape 1 : Utiliser l'algorithme de Hongrois (Hungarian algorithm) [Winston et Goldberg, 2004] pour réassigner les étiquettes des partitions d'entrée avec la partition de référence sélectionnée, comme l'un des membres de l'ensemble.

Les étapes de l'algorithme Hongrois sont données par :

Soit en entrée A la matrice qui correspond à l'affectation initiale.

Étape 1 :
Trouvez l'élément minimum dans chaque ligne de la matrice. Construisez une nouvelle matrice en soustrayant de chaque ligne de son minimum coresspondant. Pour cette nouvelle matrice, trouvez le dans chaque colonne. Construisez une nouvelle matrice en soustrayant à chaque coût le minimum de sa colonne.

Étape 2 :
Tracez le minimum de lignes (horizontales, verticales ou les deux) nécessaires pour couvrir tous les zéros de la nouvelle matrice.
Si le nombre de lignes couvrantes = la dimension de la matrice, alors une solution optimale est disponible parmi les zéros couverts dans la matrice.
Sinon, passez à **l'étape 3**.

Étape 3 :
Trouvez la valeur minimale de toutes les entrées de la nouvelle matrice ne se trouvent pas sur les lignes couvrantes.
Ensuite soustraire cette valeur de toutes les entrées qui se trouvent sur les lignes couvrantes sauf les entrées aux intersections des lignes et ajoutez cette valeur à toutes les entrées se trouvant aux intersections des lignes couvrantes.

Retournez à **l'étape 2**.

Etape 2 : Appliquer le vote majoritaire sur les partitions d'entrée réaffectées pour produire l'étiquette de cluster de la partition de consensus final [Yang, 2017].

Si on prend l'exemple de trois résultats de clustering basés sur les mêmes données d'entrée [Yang, 2017]. Le tableau 3.3 montre leurs vecteurs d'étiquettes et nous appliquons l'algorithme Hongrois avec une partition de référence $P^{'} = P_1(1,1,1,2,2,3,3)$.

	P_1	P_2	P_3
X_1	1	2	3
X_2	1	2	3
X_3	1	2	2
X_4	2	3	2
X_5	2	3	1
X_6	3	1	1
X_7	3	1	1

TABLE 3.3 – *Exemple d'ensemble de clustering.*

Après un réétiquetage, les résultats des étiquettes ré-attribués, basés sur les trois partitions d'entrée, sont présentés dans le tableau 3.4. Ensuite, le vote majoritaire peut être simplement appliqué pour déterminer l'appartenance à un cluster pour chaque élément, où l'étiquette de la partition de consensus est attribuée par la majorité des partitions d'entrée. L'étiquette de la partition de consensus final est **(1,1,1,2,2,3,3)**.

	P_1	P_2	P_3
X_1	1	1	1
X_2	1	1	1
X_3	1	1	2
X_4	2	2	2
X_5	2	2	3
X_6	3	3	3
X_7	3	3	3

TABLE 3.4 – *Relabeling*

La fonction de consensus de notre MCS peut également être décrite par le pseudo-code suivant :

Entrée :

- K (nombre de clusters obtenus pour toutes les partitions d'entrée)
- N (nombre d'objets)
- Un ensemble de partitions d'entrée $P_{k-means}, P_{KHM}, P_{BKM}, P_{SOM},$
- Une partition de référence $P^{'}$ = P selon l'indice WB
- Algorithme hongrois Hung

Pour i = 1 to 4
Réaffecter l'étiquette de la partition d'entrée : $P_i^{'} = Hang(P^{'}, P_i)$
Fin pour
$P^{'} = \left\{ P_i^{'} \right\}_i^4$
Pour i = 1 to 4
Pour n = 1 to N
Pour k = 1 to K

$$H_i^{n,k} = \begin{cases} 1 & si\ la \text{ donneé } n\ est \text{ assigneé } au\ cluster\ j\ dans\ P_i^{'} \\ 0 & Sinon \end{cases}$$

Fin pour
Fin pour
Fin pour
Pour n = 1 to N
$P_{combination}(x_n) = \arg\max_k \sum_i^4 w_i H_i^{n,k}$ $où$ $w_i = 1/4, \forall$i,
Fin pour

Sortie :
Le clustering final $P_{combination}$.

Conclusion

Dans ce chapitre, nous avons abordé le processus d'ensemble de clustering, les techniques de générations d'ensembles et les différentes fonctions de consensus.

L'approche par ensemble consiste à créer un résultat de clustering appelé consensus à partir d'un ensemble de résultats d'algorithmes de clustering. Cette approche comporte principalement deux aspects. La premier est la création de résultats de chaque clustering (différents algorithmes, différentes initialisations,etc.). La seconde est de définir une fonction qui permet de trouver la partition consensuelle finale.

Ensuite, nous avons présenté l'architecture de notre approche à savoir notre MCS (multiple classifieur système), les différentes étapes et le pseudo - code de la fonction de consensus utilisée.

Dans le chapitre suivant nous allons donner les différents résultats et discussions de nos expérimentations sur les différents jeux de données et sur les données de télédétection.

Validations expérimentales 4

Sommaire

Dans ce chapitre, nous présentons les différentes données utilisées dans notre travail ainsi que les différents résultats et discussions.

4.1 Introduction

Pour s'assurer de la validité et de la pertinence notre approche, plusieurs expériences ont été réalisées sur différents jeux de données, à savoir des données synthétiques, des images satellitaires synthétiqes et des images multispectrales. Tous les résultats des expériences ont été obtenus à l'aide du logiciel MATLAB et sont exécutés sur un PC avec processeur Intel Core i3-4000 M CPU 2.40 GHZ avec 4 GO de RAM.

4.2 Expérimentation sur des données artificielles

Dans cette section, nous proposons d'évaluer en premier lieu la qualité de l'indice WB, en second l'efficacité de l'approche proposée et en troisième la partition de référence par l'indice WB. Pour y parvenir, nous avons choisi quatorze ensembles de données artificielles différents (Clustering basic benchmark), à savoir :

- L'ensemble de données (2D) S1 -S4 qui se compose de 5000 points représentant 15 clusters et la même distribution gaussienne avec un chevauchement croissant entre les clusters.
- Les ensembles de données synthétiques (2D) A1, A2 et A3 présentant un nombre différent de clusters.
- La donnée synthétique Unbalance (2D) avec N=6500 vecteurs et k=8 clusters gaussiens.
- Les données ensembles de formes (Shape sets 2D) : Compound, D31 et R15.
- Les données UCI : Iris, Glass.
- Les données de dimension :dim032 (32D).

L'intégralité des données (datasets) peut être obtenue à partir de la page web du SIPU [http://cs.uef.fi/sipu/datasets]. Le tableau 4.1 résume les caractéristiques des données utlisées et la figure 4.1 les schématisent.

Dataset	Taille	Dimension	Nombre de Cluster
S	5000	2	15
A	3000 - 7500	2	20 - 50
Dim032	1024	32	16
Unbalance	6500	2	8
Shape sets	399 - 3100	2	6 - 31
UCI datasets	150 - 214	4 -9	3 - 6

Table 4.1 – *Données artificielles (Datasets)*

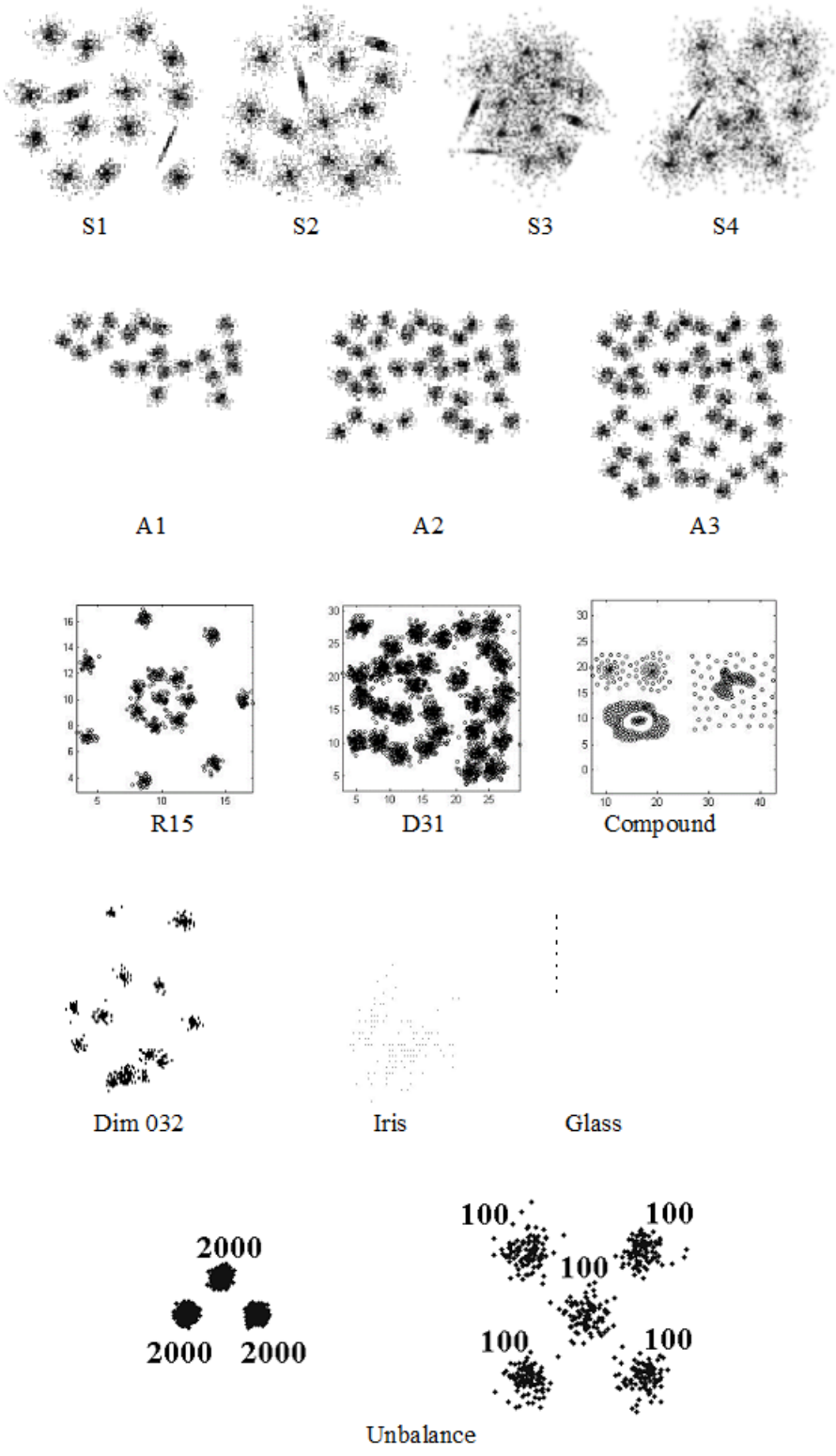


FIGURE 4.1 – *Données artificielles.*

4.2.1 Étude comparative entre les indices de validité pour obtenir le cluster optimal

Nous proposons une comparaison entre plusieurs indices de validité en les combinant avec l'algorithme k-means afin de déterminer le nombre optimal de clusters (tableau 4.2). La méthode consiste à faire varier le nombre de clusters dans un intervalle prédéfini $[k_{min}, k_{max}]$ ($K_{min} = 2$ et K_{max} est donné par la règle empirique $\sqrt{\frac{N}{2}}$ [Kodinariya et Makwana, 2013]) et à extraire les meilleures valeurs des indices représentant le nombre final de clusters. Les indices comparés sont l'indice DB, l'indice WB, l'indice BIC, l'indice IS et l'indice WSJI (voir section 2.8.1 pour une description détaillée de ces indices).

La figure 4.2 illustre les valeurs de l'indice WB sur un intervalle du nombre de clusters $k_{min} = 2$ et $k_{max} = 50$ pour les ensembles de données S1, S2, S3 et S4.

Nous pouvons constater que la valeur minmale de l'indice WB indique le bon nombre de clusters à savoir 15 clusters pour les données S1- S4.

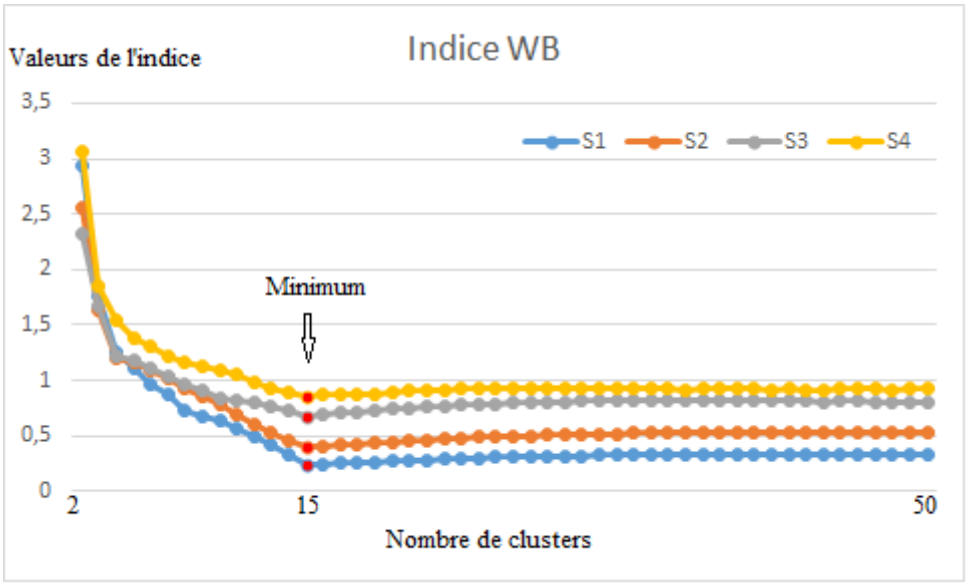


FIGURE 4.2 – *Les valeurs de l'indice WB en fonction du nombre de clusters pour les ensembles de données S1-S4.*

D'après le tableau 4.2, nous pouvons voir que les indices WB, DB, IS retournent le nombre exact de clusters pour les ensembles S1, S2, S3, S4, A1 et Dim032, de plus l'indice WB retourne le nombre exact pour les ensembles Unbalance, D31 et R15 et la valeur proche pour les autres ensembles de données. Par contre l'indice BIC retourne la valeur exact seulement pour l'ensemble Unbalance et l'indice WSJI pour les deux ensembles S1 et S3.

Les résultats obtenus montrent l'efficacité de l'indice WB à fournir un nombre optimal de clusters par rapport aux autres indices.

Données	**Nombre de cluster**	**Dimension**	**Taille**	**Nombre de cluster obtenu**				
				DB	**WB**	**BIC**	**IS**	**WSJI**
S1	15	2	5000	**15**	**15**	50	**15**	**15**
S2								13
S3								**15**
S4								20
A1	20	2	3000	**20**	**20**	39	**20**	19
A2	35		5250	34	36	50	36	33
A3	50		7500	47	53	61	47	47
Unbalance	8	2	6500	4	**8**	**8**	4	2
Compound	6	2	399	2	12	14	2	2
D31	31		3100	30	**31**	39	30	30
R15	15		600	8	**15**	17	8	8
Iris	3	4	150	2	6	9	2	2
Glass	6	9	214	7	7	10	11	4
Dim032	16	32	1024	**16**	**16**	17	**16**	15

TABLE 4.2 – *Comparaison des résultats des indices de validité*

4.2.2 Résultats de notre MCS

Nous avons appliqué notre approche, comme indiqué précédemment, pour les quatorze ensemble de données synthétiques, en faisant varier à chaque fois le nombre de clusters K pour chaque algorithme, et la valeur de l'indice WB est calculée pour les différents K. Les résultats rapportés dans les tableaux 4.3 et 4.4 illustrent les meilleures valeurs obtenues et le nombre correspondant de cluster K optimal.

Nous pouvons voir à partir des tableaux 4.3 et 4.4 que l'indice WB retourne le nombre exact de clusters pour les datasets S1, S2, S3, S4, A1, A2, A3, Unbalance, D31, R15 et Dim032 pour la plupart des algorithmes de clustering et se rapproche de la valeur optimale pour les datasets Coumpound, Iris et Glass.

Les résultats des expériences avec divers les algorithmes de clustering et les datasets montrent que l'indice WB fournit une estimation fiable du nombre de clusters, ce qui suggère l'efficacité de la validité des indices internes de clustering. Contrairement à la F-mesure obtenue en utilisant chaque algorithme de clustering seul, nous avons obtenu une F-mesure élevée en utilisant la méthode de combinaison. Ces résultats montrent que le MCS peut augmenter la précision du clustering [Chaouche Ramdane et al., 2022].

Dataset	Approche	K optimale	Indice WB	F-mesure
S- Sets Données Synthétiques d=2 (N = 5000 vecteurs, K = 15 clusters gaussiens)				
S1	K- means	**15**	0.2355	0.9938
	KHM		0.2367	0.9934
	BKM		0.2355	0.9936
	SOM		0.2429	0.9936
	MCS		**0.2352**	**0.9946**
S2	K- means	**15**	0.3954	0.9696
	KHM		0.3930	0.9688
	BKM		0.3954	0.9696
	SOM		0.4192	0.9606
	MCS		**0.3950**	**0.9700**
S3	K- means	**15**	0.6770	0.8550
	KHM		0.7633	0.8456
	BKM		0.6770	0.8568
	SOM		0.8040	0.8182
	MCS		**0.6637**	**0.8600**
S4	K- means	**15**	0.8607	0.7968
	KHM		0.9090	0.7398
	BKM		0.8607	0.7974
	SOM		1.0356	0.7677
	MCS		**0.8604**	**0.8000**
A - Sets **Données synthètiques d=2 (150 vecteurs par cluster)**				
A1 **(N = 3000, 20 clusters)**	K- means	**20**	0.2268	0.9987
	KHM	**20**	0.2888	0.9257
	BKM	**20**	0.2268	0.9987
	SOM	21	0.2630	0.9755
	MCS	**20**	**0.2260**	**0.9990**
A2 **(N = 5250, 35 clusters)**	K- means	36	0.3026	0.9862
	KHM	39	0.3867	0.8731
	BKM	**35**	0.2948	0.9989
	SOM	36	0.3109	0.9723
	MCS	**35**	**0.2940**	**0.9990**
A3 **(N = 7500, 50 clusters)**	K- means	53	0.3413	0.9400
	KHM	54	0.3489	0.9040
	BKM	**50**	0.3097	0.9972
	SOM	51	0.3184	0.9840
	MCS	**50**	**0.3093**	**0.9991**
Unbalance **(Données synthètiques d=2)**				
Unbalance **(N = 6500, 8 Gaussian clusters)**	K- means	**8**	**0.0335**	**1**
	KHM	9	0.0373	0.8495
	BKM	**8**	**0.0335**	**1**
	SOM	**8**	0.0543	0.8917
	MCS	**8**	**0.0335**	**1**

TABLE 4.3 – *Résultats sur les données synthétiques (S, A, Unbalance)*

Dataset	Approche	K optimale	Indice WB	F - mesure
Shape sets (Données synthètiques d = 2)				
Compound (N = 399, 6 clusters)	K - means	**11**	0.4836	0.4812
	KHM	**11**	0.4845	0.4912
	BKM	**11**	0.4874	0.4887
	SOM	12	0.4909	0.4761
	MCS	**11**	**0.4665**	**0.5438**
D31 (N = 3100, 31 clusters)	K - means	**31**	0.3832	0.9335
	KHM	33	0.4317	0.9000
	BKM	**31**	**0.3459**	0.9761
	SOM	32	0.3695	0.9529
	MCS	**31**	**0.3459**	**0.9768**
R15 (N = 600, 15 clusters)	K - means	**15**	**0.1287**	**0.9967**
	KHM	16	0.1423	0.8874
	BKM	**15**	**0.1287**	**0.9967**
	SOM	12	0.3891	0.5300
	MCS	**15**	**0.1287**	**0.9967**
UCI data sets				
Iris (N = 150, 3 clusters)	K - means	6	0.3639	0.5133
	KHM	**4**	0.3585	0.7000
	BKM	5	0.3671	0.6733
	SOM	8	0.3810	0.2266
	MCS	**4**	**0.3118**	**0.7067**
Glass (N= 214, 6 clusters)	K - means	**7**	1.9474	0.5377
	KHM		2.1751	0.5327
	BKM		2.1679	0.5359
	SOM		2.3012	0.5238
	MCS		**1.9471**	**0.5472**
Dim - sets (grande dimensions)				
Dim 032 (N = 1024, 16 clusters gaussians)	K - means	**16**	**0.0389**	**1**
	KHM	**16**	0.0410	**1**
	BKM	**16**	0.0389	**1**
	SOM	14	0.0951	0.5352
	MCS	**16**	**0.0389**	**1**

TABLE 4.4 – *Résultats sur les données synthètiques (Shape sets, UCI datasets, Dim-sets)*

4.2.3 Comparaison avec la partition médiane

Afin de valider le choix de la partition de référence par l'indice WB optimale, nous avons appliqué l'approche par la partition médiane à l'ensemble des partitions des données synthétiques.

Les résultats rapportés dans le tableau 4.5 montrent que la partition de référence résultante est pratiquement la même que la partition donnée par l'indice WB, ce qui démontre la qualité de l'indice WB.

Données	**Partition de référence**	
	Approche WB optimale	**Approche médiane**
S1 **S2** **S3** **S4**	BKM / K - means	BKM
A1	BKM / K - means	BKM
A2	BKM	BKM
A3	BKM	K- means
Unbalcane	BKM / K - means	K - means
Compound	K - means	K - means
D 31	BKM	K - means
R 15	BKM / K - means	BKM
Iris	KHM	KHM
Glass	K - means	KHM
Dim 032	BKM / K - means	BKM

TABLE 4.5 – *Comparaison de la partition de référence entre les approches indice WB et partition médiane.*

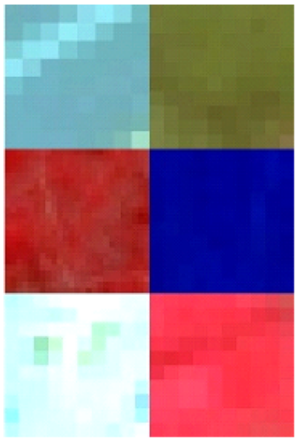
4.3 Expérimentation sur des images composites

Nous proposons dans cette section d'appliquer notre approche sur trois images satellites synthétiques présentant 6, 8 et 10 clusters (Figure 4.3). Les données, comme mentionné ci-dessus, ont été collectées à partir des images satellitaires en sélectionnant manuellement des échantillons pertinents. Ensuite, les échantillons sont redimensionnés et concaténés en une image composite finale. Cette façon de procéder, nous permet de construire des images satellites synthétiques avec des classes connues afin d'effectuer les tests.

A travers les résultats obtenus dans le tableau 4.6, nous constatons que notre MCS retourne un nombre de cluster proche du nombre de cluster des images composites avec un indice WB minimum pour l'image résultat de la combinaison par rapport à l'indice WB obtenu à l'aide des algorithmes de clustering [Chaouche Ramdane et al., 2019]. Ces résultats illustrés par les figures 4.4, 4.5 et 4.6 montrent, également, que le MCS améliore la qualité de la classification des images.

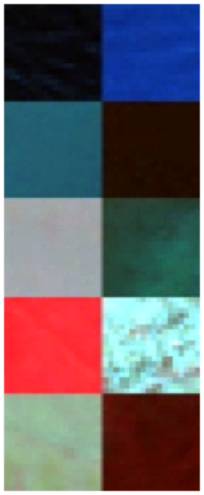
Image composite	Approche	K optimal	Indice WB
(a) **6 clusters**	K- means	7	0.0767
	KHM	8	0.0982
	BKM	7	0.1187
	SOM	8	0.1434
	MCS	**7**	**0.0759**
(b) **8 clusters**	K- means	9	0.0759
	KHM	12	0.2497
	BKM	12	0.0768
	SOM	11	0.1594
	MCS	**9**	**0.0752**
(c) **10 clusters**	K- means	12	0.0836
	KHM	16	0.1088
	BKM	16	0.0853
	SOM	15	0.1608
	MCS	**12**	**0.0790**

Table 4.6 – *Résultats sur les images composites.*



(a)

(b)



(c)

FIGURE 4.3 – *Les images composites.*

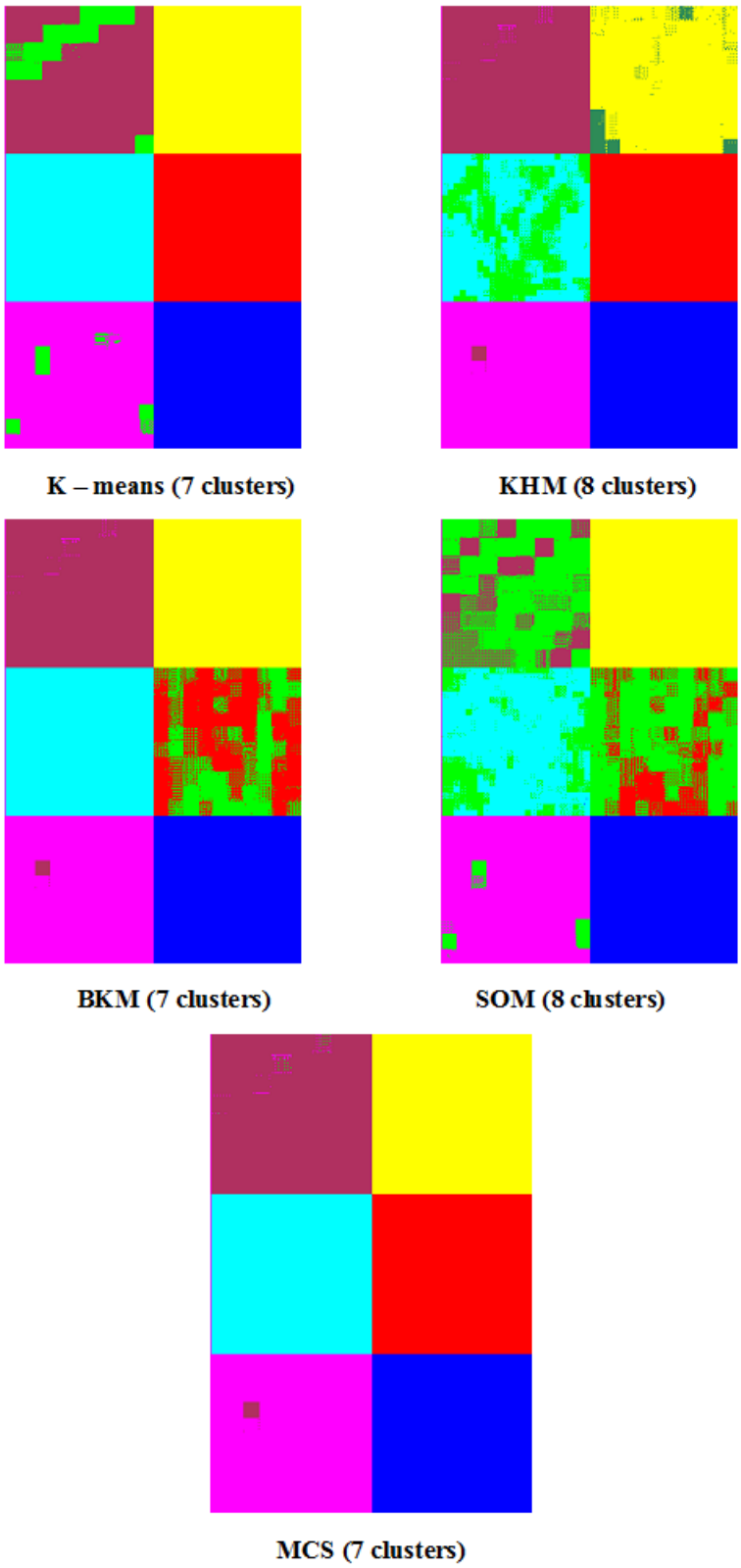


FIGURE 4.4 – *Résultat de l'approche MCS sur l'image composite (a).*

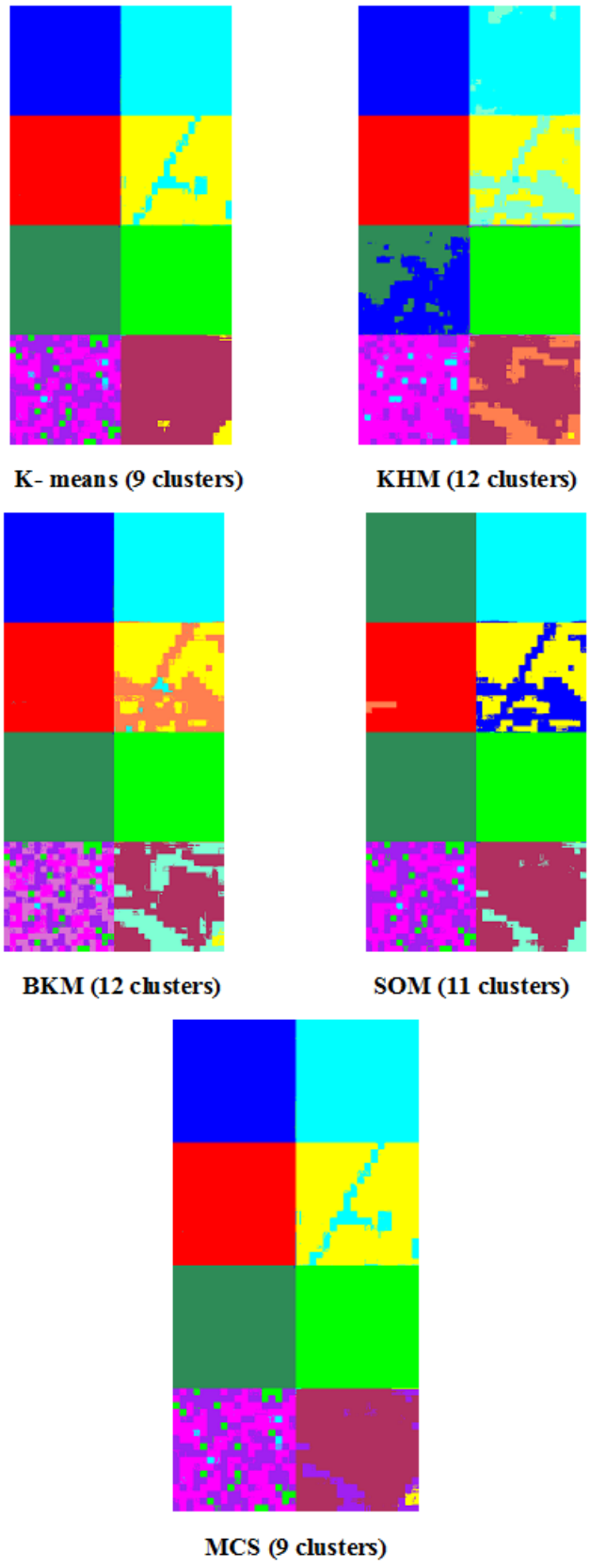


FIGURE 4.5 – *Résultat de l'approche MCS sur l'image composite (b).*

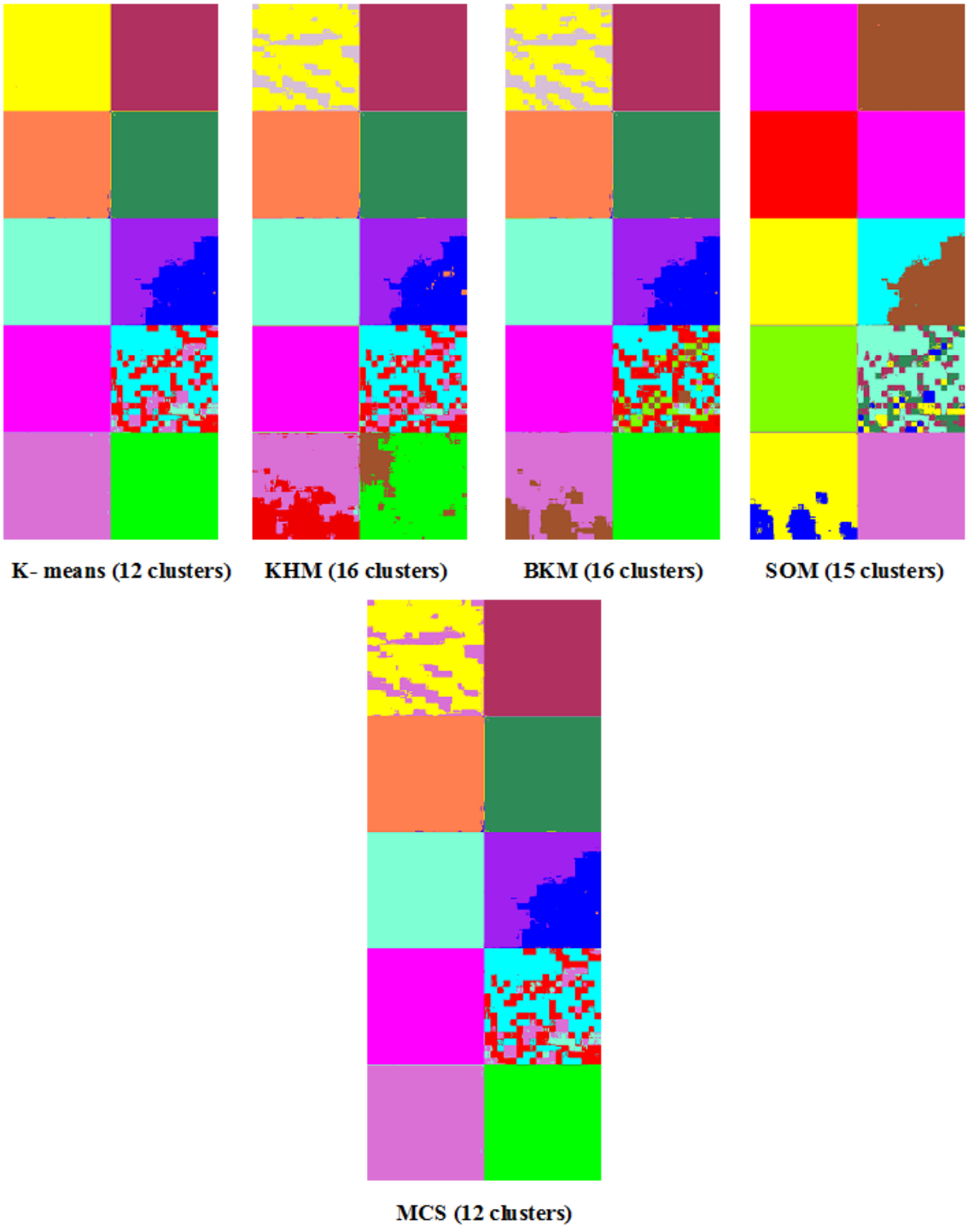


FIGURE 4.6 – *Résultat de l'approche MCS sur l'image composite (c).*

4.4 Expérimentation sur des images multispectrales

Dans cette dernière section, nous proposons d'appliquer l'approche MCS sur quatre images multispectrales, qui représentent les régions d'Oran et de Tlemcen en Algérie ; les caractéristiques de chaque image sont reportées dans le tableau 4.7. Ces régions sont relativement complexe, avec des thématiques assez diversifiées (Figure 4.7) : Urbain, Sebkha, sol, eau, végétation, forêt, relief et culture. L'image multispectrale d'Oran (scène 3) contient un tissu urbain très hétérogène. En effet, les anciennes constructions, apparaissent sur l'image avec une texture plus contrastée et un aspect moucheté, tandis que le nouveau bâti est plus homogène et moins contrasté. On y trouve également des zones à topographie relativement complexe. On est donc devant un cas relativement difficile. Ceci confirme l'intérêt d'utiliser des méthodes de clustering pour des tâches similaires.

Les résultats du clustering des quatre images multispectrales par l'approche proposée sont présentés dans les figures 4.8, 4.9, 4.10 et 4.11.

Nous avons obtenu l'indice WB optimale pour l'image résultat du multi classifieur par rapport aux images des algorithmes de clustering. Aprés une analyse visuelle des résultats avec les images originales correspondantes, les résultats obtenus semblent être satisfaisants en général. Cependant nous remarquons une certaine confusion entre les pixels d'eau et les pixels d'ombre, dans le cas de la première image (scène1) et d'autres confusions sur certaines zones non urbaines apparaissent. C'est le cas du relief (première image) et de la zone de Sebkha (troisième image) qui présente une variabilité spatiale similaire aux zones bâties [Chaouche Ramdane et al., 2022].

	Taille (m^2)	**Résolution**	**Satellite**	**Région (Ouest Algérie)**
Scène 1	512 x 512	30 m	Landsat	Oran - ville
Scène 2	400 x 400	20 m	Spot	Oran - rural
Scène 3	500 x 500	15 m	Aster	Oran - ville
Scène 4	400 x 400	2.5 m	Alsat 2a	Tlemcen

Table 4.7 – *Caractéristiques des images multispectrales.*

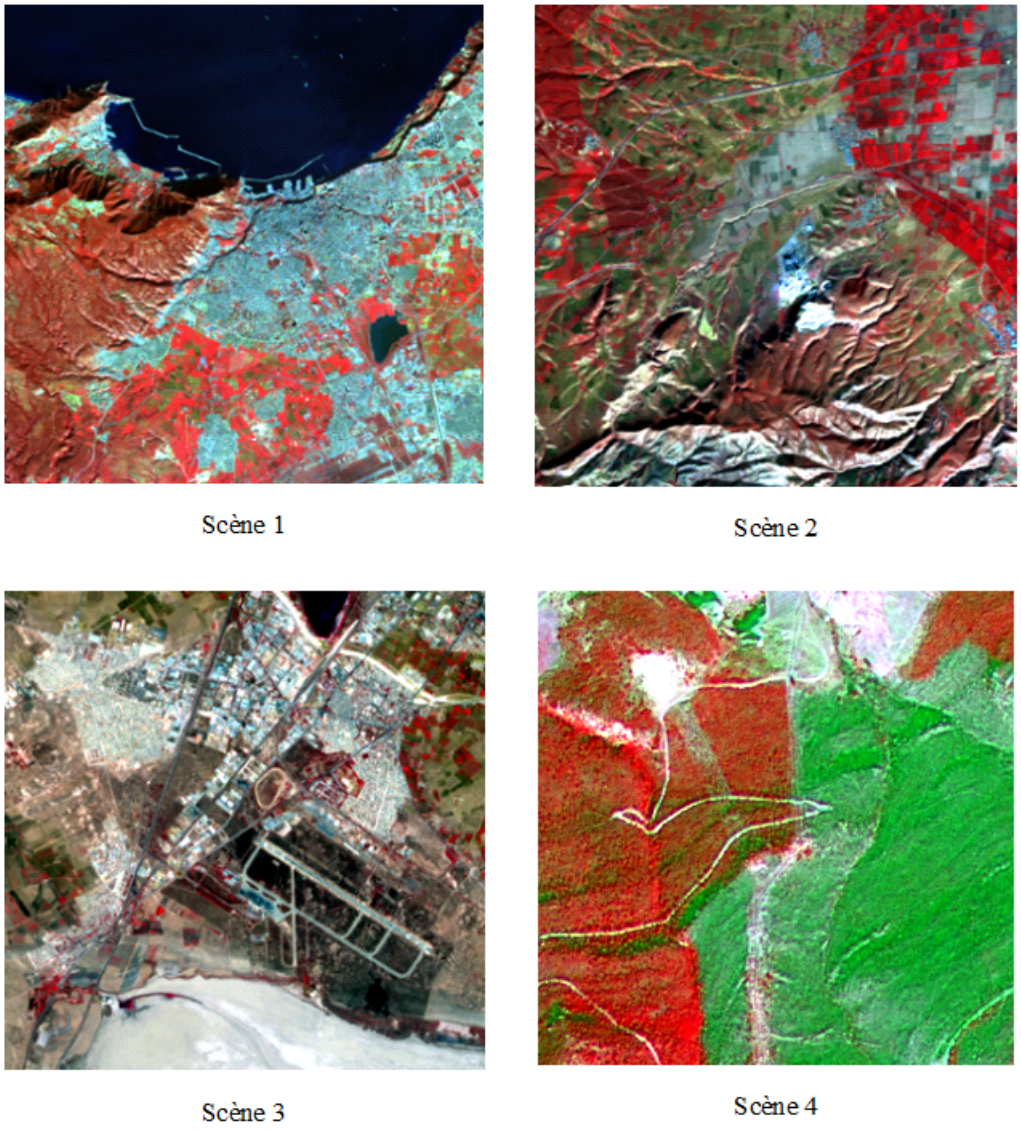


FIGURE 4.7 – *Les images multispectrales.*

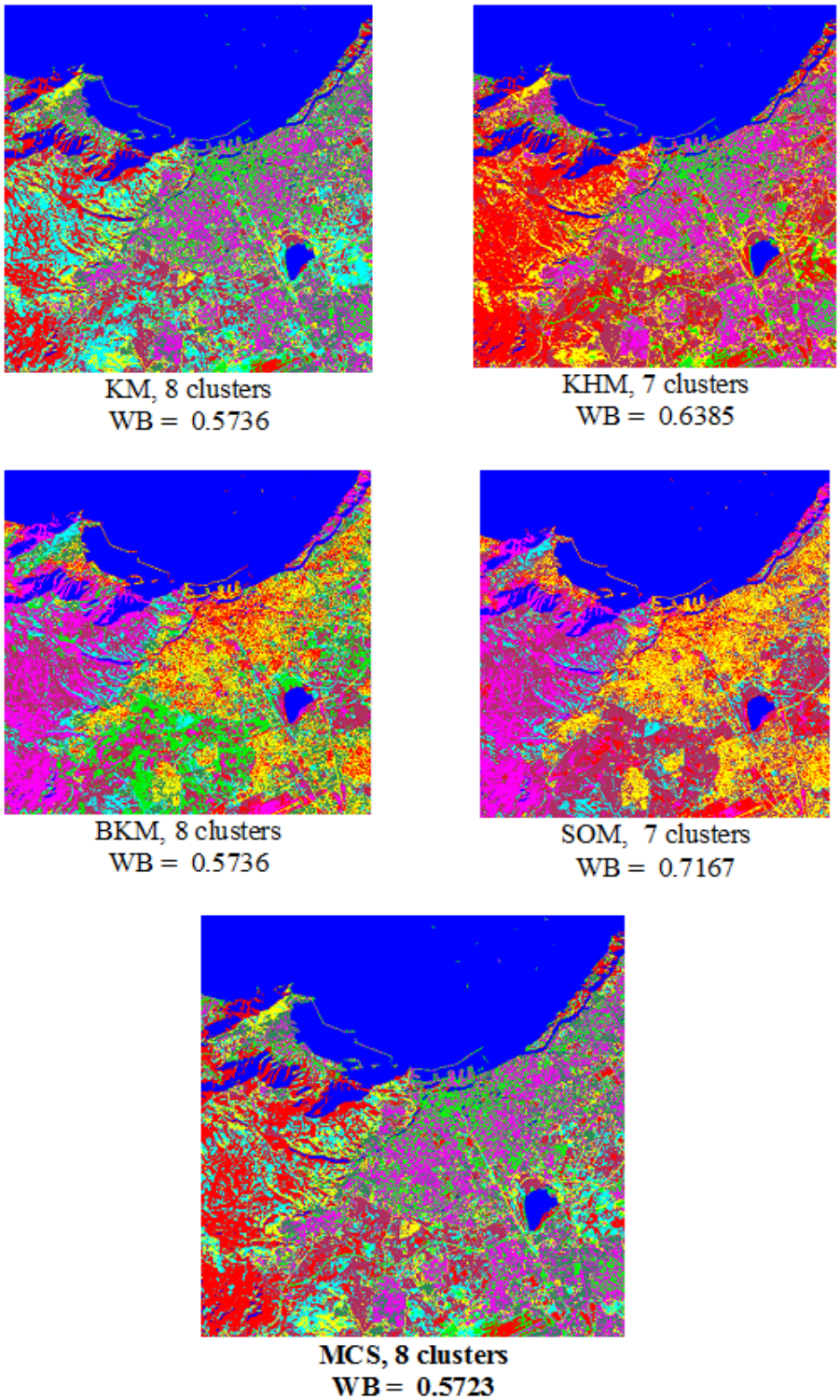


FIGURE 4.8 – *Résultat de l'approche MCS sur la scène 1.*

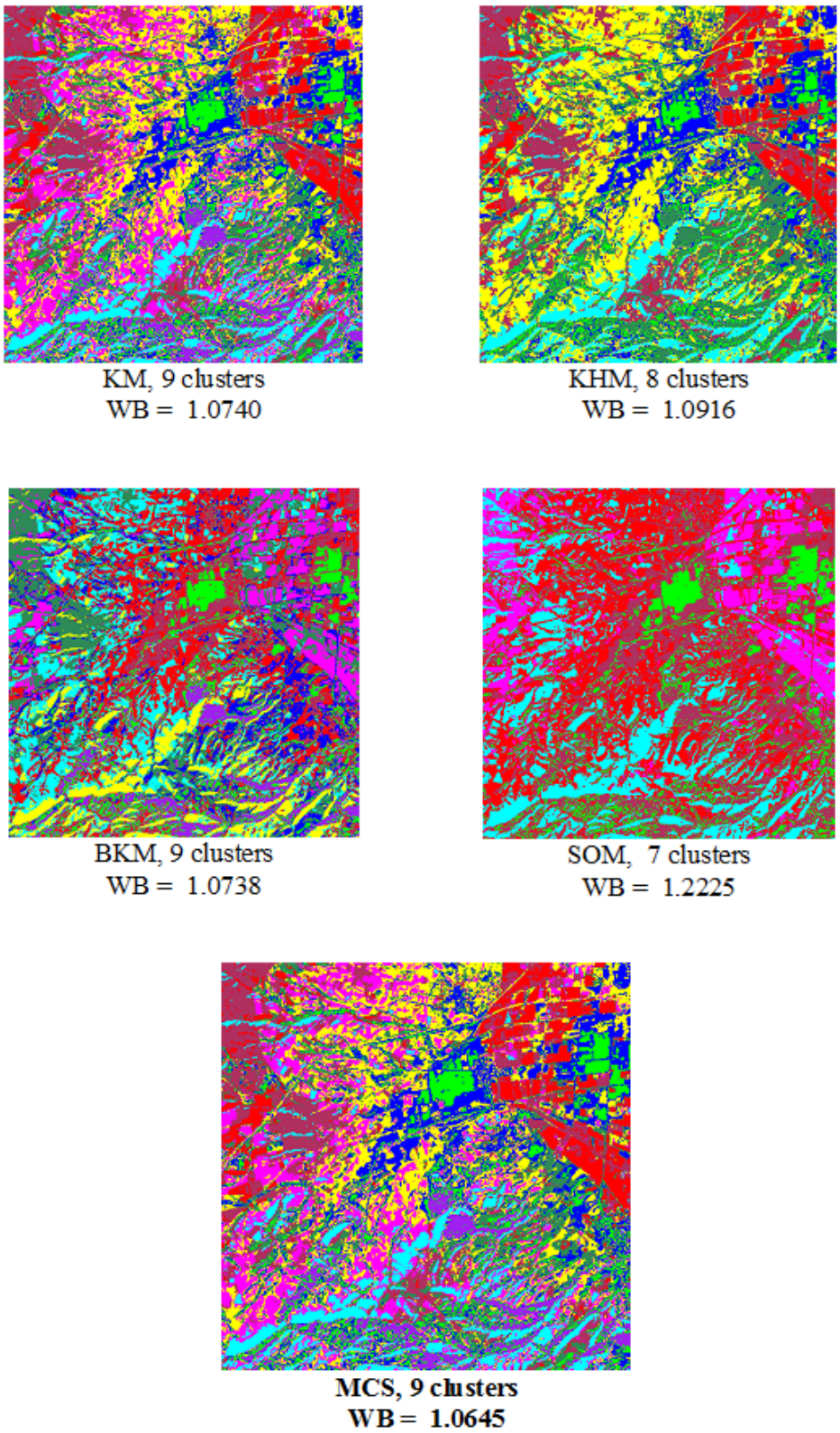


FIGURE 4.9 – *Résultat de l'approche MCS sur la scène 2.*

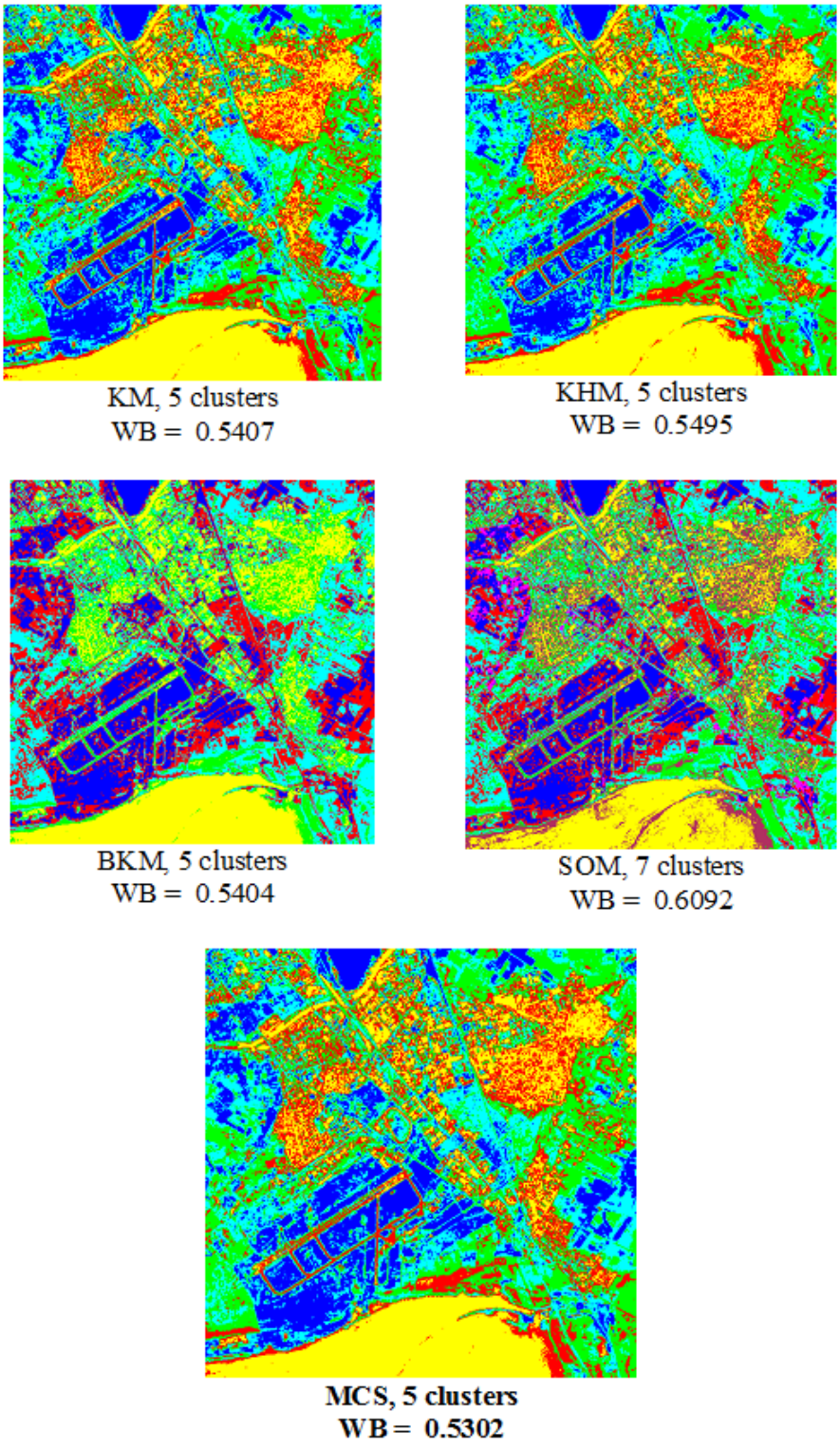


FIGURE 4.10 – *Résultat de l'approche MCS sur la scène 3.*

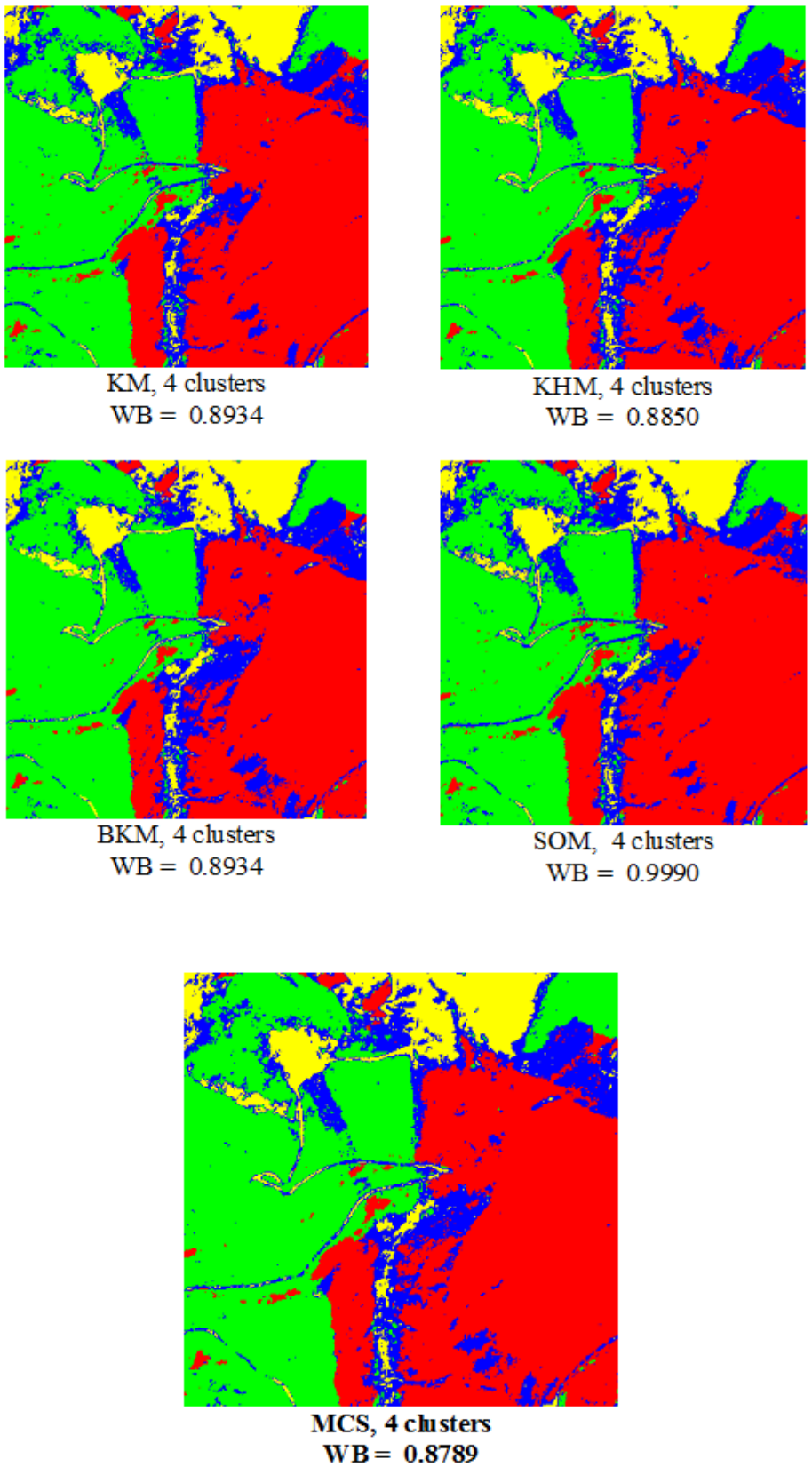


FIGURE 4.11 – *Résultat de l'approche MCS sur la scène 4.*

4.4.1 Temps d'exécution du MCS

Malgré les bons résultats, notre MCS est gourmand en temps car de nombreux calculs sont nécessaires à son exécution à savoir les calculs pour l'étape de génération des quatre partitions et l'étape de combinaison de partitions (Figure 4.12) sachant en plus que le temps d'exécution de la partie extraction de l'indice WB optimale d'une partition d'un clustering prenait plus de 30 minutes.

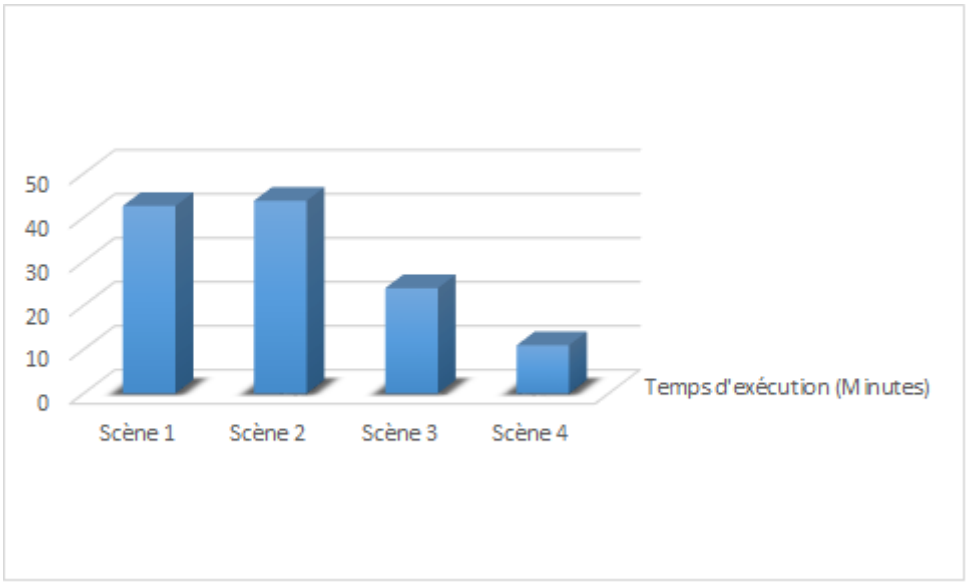


FIGURE 4.12 – *Temps d'exécution du MCS pour les images multispectrales.*

4.4.2 Classification par maximum de vraisemblance

Pour valider nos résultats, une classification supervisée par maximum de vraisemblance [Richards, 1993] a été effectuée. Ce processus était basé sur la scène 2 et la scène 3 .

L'image de la scène 2 est classée en neuf classes qui représentent les classes suivantes : relief #1 (R1), relief #2 (R2), forêt (F), végétation #1 (V1), végétation #2 (V2), sol nu, terres agricoles (T), urbain (U), et nuages (N).

Le résultat est illustré dans la figure 4.13. Le test générée par cette approche a montré une précision de classification de 91.7480 % avec une statistique de Kappa de 0.9052. Le résultat de la matrice de confusion est présenté dans le tableau 4.8 .

Notre approche produit une image classée en neuf classes, soit le même résultat que la classification supervisée ce qui révèle la force de notre MCS.

Quand à l'image de la scène 3 est classée en six classes qui représentent les classes urbain #1 (U1), urbain #2 (U2), végétation #1 (V1), végétation #2 (V2), sol et Sebkha (Se), le résultat est montré dans la figure 4.14.

De plus, le résultat de la matrice de confusion pour la classification supervisée est reporté dans le tableau 4.9, et la précision de test générée par cette approche a montré une précision de classification de 79,36 % avec une statistique de Kappa de 0,74.

Une comparaison visuelle montre que les résultats sont presque similaires. Notre approche produit une image classée en cinq classes, soit une classe de moins que la classification supervisée. Cette dernière qui représente la classe Sebkha, est une région qui a été confondue avec une classe urbaine en raison des valeurs spectrales similaires entre les classes. Ainsi, les deux classes urbaines (Urbain #1 et Urbain #2) regroupent 83798 points pour la classification supervisée alors que dans les résultats du MCS, 89034 points sont regroupés sous ces classes.

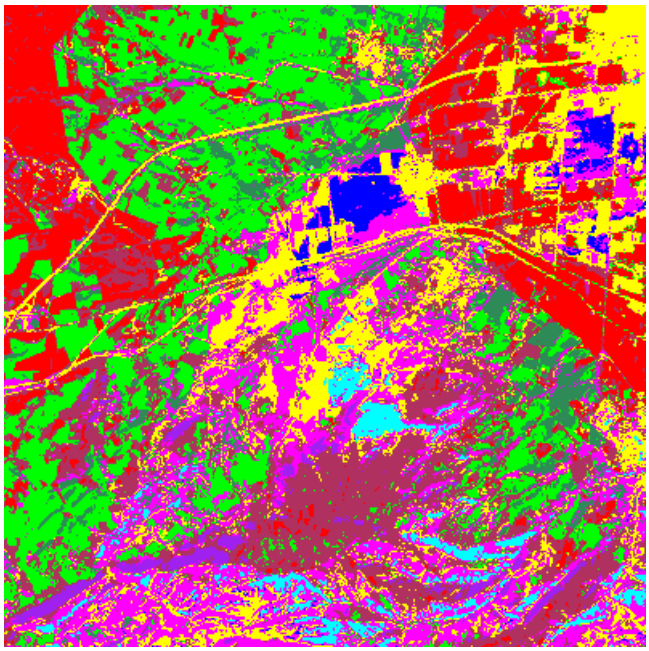


FIGURE 4.13 – *Classification supervisée sur la scène 2.*

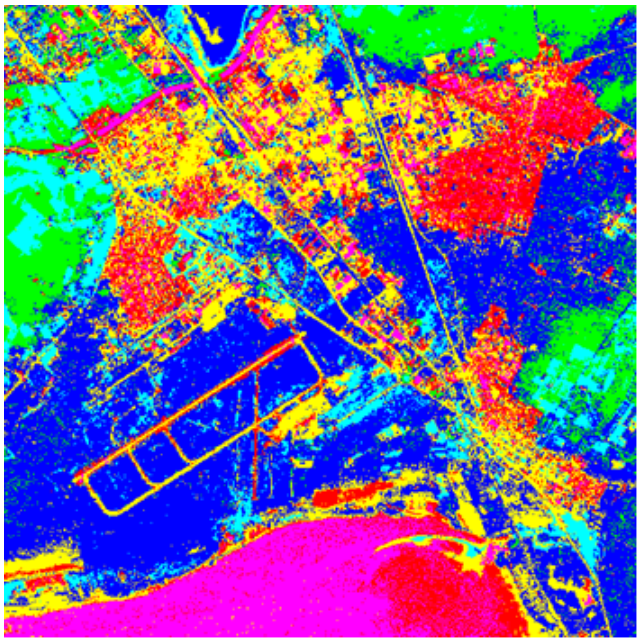


FIGURE 4.14 – *Classification supervisée sur la scène 3.*

Données classifiées	Données de référence										
Classes d'occupation du sol	**R1**	**R2**	**F**	**V1**	**V2**	**T**	**Sol**	**U**	**N**	**Total lignes**	**Précision de l'utilisateur %**
Relief 1	301	4	0	0	0	0	0	0	0	305	98.68
Relief 2	6	369	0	23	2	0	10	0	0	410	90
Forêt	0	0	128	1	3	11	0	9	6	158	81.01
Végétation 1	0	5	1	445	3	0	0	0	0	454	98.01
Végétation 2	0	0	0	38	190	0	0	0	0	228	83.33
Terres agricoles	0	0	9	0	0	339	0	4	0	352	96.30
Sol nu	0	4	0	0	0	0	113	0	0	117	96.58
Urbain	0	2	34	0	0	3	0	194	0	233	83.26
Nuages	0	0	24	0	0	0	0	1	178	203	87.68
Total colonne	307	384	196	507	198	353	123	208	184	2460	**Précision globale**
Précision du générateur	98.05	96.09	65.31	87.77	95.96	96.03	91.87	93.27	96.74		**91.7480 %**

TABLE 4.8 – *Matrice de confusion de la scène 2 de la classification supervisée.*

Données classifiées	Données de référence							
Classes d'occupation du sol	**U1**	**U2**	**V1**	**V2**	**Sol**	**Se**	**Total lignes**	**Précision de l'utilisateur %**
Urbain 1	172	18	0	0	0	170	360	47.79
Urbain 2	16	126	0	0	19	0	161	78.26
Végétation 1	0	0	168	3	0	0	171	98.24
Végétation 2	0	1	0	137	0	0	138	99.27
Sol	9	2	7	0	148	0	166	89.18
Sebkha	13	2	0	0	0	249	264	94.32
Total colonne	210	149	175	140	167	419	1260	**Précision globale**
Précision du générateur	81.90	84.56	96	97.86	88.62	59.43		**79.36 %**

TABLE 4.9 – *Matrice de confusion de la scène 3 de la classification supervisée.*

Conclusion

Nous avons évalué notre MCS sur quatorze ensembles de données synthétiques, trois données composites et quatre images multispectrales en utilisant les algorithmes K- means, KHM, BKM et SOM pour le clustering des ensembles de données. Ensuite, en fonction de la valeur minimale de l'indice WB, le nombre final de clusters a été retenu. Le processus se termine par l'application de l'étape de la combinaison des partitions. Finalement, l'indice WB donne des résultats positifs, pour la plupart des ensembles de données synthétiques, il fournit le nombre exact de clusters pour toutes les approches de clustering, avec une F-mesure élevée pour le résultat combiné par rapport aux F-mesures des méthodes de clustering.

La comparaison visuelle montre que les résultats du MCS sur les images de télédétection sont bons et satisfaisants.

Afin de valider le choix de la partition de référence par l'indice WB optimale, une comparaison avec l'approche de la partition médiane à l'ensemble des partitions des données synthétiques est effectuée. Les résultats obtenus démontrent la qualité de l'indice WB.

Il est à noter que le calcul de notre MCS nécessite beaucoup d'effort de calcul.

Pour conclure, nous avons évalué notre approche par une comparaison avec une méthode de classification supervisée (maximum de vraisemblance). Les résultats obtenus sont bons, ce qui démontre la qualité de notre approche. Ainsi, le MCS a montré un grand potentiel pour améliorer la classification non supervisée des images satellitaires.

Mes contributions scienfifiques

- Chaouche Ramdane, L., Mahi, H. et Bessaid, A. Combination of different clustering algorithms. In : The 6th International Conference on Image and Signal Processing and Their Applications ISPA 2019, Algeria (2019).
  DOI :10.1109/ISPA48434.2019.8966871

- Chaouche Ramdane, L., Mahi, H., El Habib Daho, M. et Lazouni, M.A. Multiple classifier system for remotely sensed data clustering. IET Image Process. 1(16), pages 252-260, (2022). https ://doi.org/10.1049/ipr2.12349

Conclusion générale

Sommaire

Si microscope et télescope évoquent les grandes percées scientifiques vers l'infiniment petit et l'infiniment grand, nous avons besoin aujourd'hui d'un macroscope pour explorer l'infiniment complexe.

-J. de ROSNAY-

Synthèse

Ces dernières années, la résolution spatiale des données satellitaires s'est améliorée grâce aux progrès de la technologie satellitaire, et l'on peut s'attendre à l'avenir à l'acquisition d'informations encore plus détaillées sur la surface de la terre. Cependant, diverses complications sont encore associées à la récupération et à la classification des informations sur la surface terrestre à partir de données satellitaires à très haute résolution. La classification est une tâche importante en reconnaissance des formes, ce qui est la raison principale pour laquelle les dernières décennies ont vu un grand nombre de projets de recherche consacrés aux méthodes de classification.

La limitation des méthodes de classification supervisée, quel que soit le classifieur, réside dans la nécessité d'utiliser des échantillons et donc une vue restrictive et non objective de la réalité. Il est même parfois impossible d'effectuer une classification supervisée faute de disponibilité de vérités terrain (année passée, zone inconnue ou inaccessible).

Pour cette raison, les systèmes à classificateurs multiples constituent une orientation importante dans l'apprentissage automatique et la reconnaissance des formes. En effet, la combinaison de classificateurs est désormais un domaine de recherche actif, connu sous différents noms dans la littérature, tels que le mélange d'experts, l'apprentissage par comité ou les méthodes d'ensemble et bien que le MCS ait un grand nombre d'applications dans le traitement des données de télédétection, il convient de mentionner que la plupart des travaux impliquent une classification supervisée ou semi-supervisée, alors que seules quelques études ont été consacrées sur des cas non supervisés.

C'est dans ce contexte que s'est porté notre contribution. L'objectif principal de cette thèse était d'évaluer l'efficacité d'un système de classification multiple (MCS), qui combinent les résultats de différents classificateurs non supervisés.

En effet, l'étape fondamentale d'une méthode de clustering d'ensemble est la fonction de consensus. Elle combine de nombreux partitions pour produire un résultat de clustering amélioré par rapport aux partitions individuels de l'ensemble. Il existe deux approches principales de la fonction de consensus : la co-occurrence et la partition médiane. Dans la première approche, on trouve les méthodes basées sur le relabeling et le vote, la matrice de co-association et les graphes.

Notre approche est basée sur le ré-étiquetage et le vote. cette dernière tente de résoudre le problème de correspondance, puis un simple vote peut être appliqué pour affecter les objets dans les clusters afin de déterminer la partition consensuelle finale.

Contributions

Au cours de cette thèse, nous avons évalué notre MCS sur plusieurs ensembles de données synthétiques, trois images composites et quatre images multispectrales issues de quatre capteurs différents en utilisant les algorithmes K-means, KHM, BKM et SOM pour le clustering des ensembles de données. Ensuite, en fonction de la valeur minimale de l'indice de validité interne du clustering WB, le nombre final de clusters a été retenu et le meilleur clustering qui fait office de référence est obtenu. Le processus se termine par l'application de l'étape de la combinaison des partitions en utlisant l'approche relabeling et vote.

Grâce à cette approche, nous avons pu, obtenir des bons résultats. Pour la plupart des ensembles de données synthétiques,l'indice WB fournit le nombre exact de clusters pour toutes les approches de clustering, avec une précision de F-mesure élevée pour le résultat combiné par rapport aux F-mesures des quatre algorithmes de clustering.

Pour évaluer le choix de la partition de référence par l'indice WB optimale, une comparaison avec l'approche de la partition médiane à l'ensemble des partitions des données synthétiques est effectuée. Les résultats obtenus montrent l'efficacité de l'indice WB.

La comparaison visuelle montre que les résultats des images de télédétection sont bons et satisfaisants.

Afin de situer l'approche développée, une comparaison avec une méthode de classification supervisée la plus utilisée en classification d'images satellitaires à savoir le maximum de vraisemblance est effectuée sur deux images multispectrales. Les résultats obtenus ont mis en évidence la qualité de notre approche.

Ainsi, le MCS a montré un grand potentiel pour améliorer la classification non supervisée des données de télédétection, cependant ce processus est coûteux en temps car de nombreux calculs sont nécessaires à son exécution.

Perspectives

Le travail présenté dans cette thèse, peut avoir un impact sur la suite des travaux de recherches à entreprendre dans l'avenir. L'ensemble de clustering nous a permis d'obtenir des résultats très concluants qui méritent d'être consolidés en prospectant d'autres horizons, on peut citer notamment :

- Le test avec d'autres indices de validité de cluster et d'autres algorithmes de clustering tels que le clustering basé sur la densité (DENCLUE, DBSCAN, OPTICS), le clustering basé sur la grille (STING, CLIQUE, optiGrid), la technique de clustering basée sur les algorithmes génétiques, le clustering GMM (Gaussian mixture models) et etc.

- Le test et la validation de cette approche sur d'autres types d'images optiques (résolutions différentes, différentes régions), afin d'étudier l'effet du changement de la résolution, ainsi que la comparaison des résultats sur d'autres régions.

- D'explorer d'autres familles d'ensemble de clustering ainsi que d'autres jeux de données.

Bibliographie

H. Akaike. A new look at the statistical model selection identification. *IEEE Transaction on Automatic Control*, 19 :719–723, 1974.

O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J.M. Perez, et I. Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46 : 243–256, 2013.

H.G. Ayad et M. S. Kamel. Cumulative voting consensus method for partitions with a variable number of clusters. *IEEE Trans. Patt. Anal. Mach. Intell.*, 30(1) :160 – 173, 2008.

H.G. Ayad et M. S. Kamel. Bagging to improve the accuracy of a clustering procedure. *Pattern Recognition*, 43(5) :1943 – 1953, 2010.

Wim H. Bakker et Al. *Principles of remote sensing*. The International Institute for Geo-Information Science and Earth Observation (ITC), Hengelosestraat 99, P.O. Box 6, 7500 AA Enschede, The Netherlands, 2004.

G. Ball et D. Hall. Isodata : a novel method of data analysis and pattern classification. *Technical report, Stanford Research Institute,Menlo Park*, 1965.

J.A. Benediktsson, J. Chanussot, et M. Fauvel. Multiple classifier systems in remote sensing : From basics to recent developments. *Mult. Classif. Syst.*, 4472 :501–512, 2007.

V. Berikov, I. Pestunov, P. Melnikov, et G. Gonzalez. Centroid-based ensemble clustering : algorithms for hyperspectral images segmentation. *Proceedings of the 9th Open German-Russian Workshop on Pattern Recognition and Image Understanding*, pages 56–59, 2014.

J.C. Bezdek, R. Ehrlich, et W. Full. Fcm : the fuzzy c-means clustering algorithm. *Comput. Geosci.*, 10(2–3) :191–203, 1984.

J. Bodersen. Multispectral lighting : A practical option for difficult industrial imaging situations. *Photonics spectra*, 2021.

G. Briem, J. Benediktsson, et J. Sveinsson. Multiple classifiers applied to multisource remote sensing data. *IEEE Trans. Geosci. Remote Sens.*, 40 : 2291–2299, 2002.

H. J. Buiten et J. G. P. W. Clevers. *Land Observation by Remote Sensing : Theory and Applications*. vol. 3 of Current Topics in Remote Sensing. Gorden and Breach, 1993.

G. Camps-Valls et L Bruzzone. *Kernel Methods for Remote Sensing Data Analysis*. John Wiley and Sons, Ltd, Publication, 2009.

L. Chaouche Ramdane, H. Mahi, et A. Bessaid. Combination of different clustering algorithms. *International Conference on Image and Signal Processing and Their Applications ISPA 2019, Algeria*, 2019.

L. Chaouche Ramdane, H. Mahi, M. El Habib Daho, et M.E.A. Lazouni. Multiple classifier system for remotely sensed data clustering. *IET Image Process*, 1(16) :252–260, 2022.

M. Chi, Q. Kun, J.A. Benediktsson, et R. Feng. Ensemble classification algorithm for hyperspectral remote sensing data. *IEEE Geosci. Remote Sens. Lett.*, 6(4) :762–766, 2009.

A. Cichocki, M. Mrup, P. Smaragdis, W. Wang, et R. Zdunek. Advances in nonnegative matrix and tensor factorization. *Computational Intelligence and Neuroscience (Hindawi Publishing Corporation)*, 2008.

J. Cocquerez et S. Philipp. *Analyse d'images : filtrage et segmentation*. Elsevier-Masson, 1995.

M.N.D. Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. De l'imprimerie royale, Paris, 1785.

T. Cover et P. Hart. Nearest neighbour pattern classification. *IEEE Transactions on Information Theory*, 13 :21–27, 1967.

D.L. Davies et D.W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2) :224 – 227, 1979.

J. Davis et M. Goadrich. The relationship between precision-recall and roc curves. *Proceedings of the 23rd International Conference on Machine learning*, page 233–240, 2006.

E. Dimitriadou, A. Weingessel, et K. Hornik. An ensemble method for clustering. *ICANN*, pages 217 –224, 2001.

E. Dimitriadou, A. Weingessel, et K. Hornik. A combination scheme for fuzzy clustering. *Int. J. Patt. Recogn. Artif. Intell.*, 16(7) :901 –912, 2002.

H.T.X. Doan et G.M. Foody. Increasing soft classification accuracy through the use of an ensemble of classifiers. *Int. J. Remote Sens.*, 28 :4609–4623, 2007.

E. Duchesnay. *Agents situés dans l'image et organisés en pyramide irrégulière. Contribution à la segmentation par une approche d'agrégation coopérative et adaptative*. PhD thesis, Université de Rennes 1, 2001.

S. Dudoit et J Fridlyand. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, 19(9) :1090 – 1099, 2003.

J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3 :32–57, 1973.

T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8) : 861–874, 2006.

X. Z. Fern et C. E. Brodley. Solving cluster ensemble problems by bipartite graph partitioning. *Proceedings of the 21st International Conference on Machine learning, ACM*, page 36, 2004.

B. Fischer et J. M. Buhmann. Bagging for path-based clustering. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 25(11) :1411 – 1415, 2003.

G.M. Foody, D.S. Boyd, et C. Sanchez-Hernandez. Mapping a specific class with an ensemble of classifiers. *Int. J. Remote Sens.*, 28 :1733–1746, 2007.

G. Forestier, C. Wemmert, et P. Gançarski. Collaborative multi-strategical clustering for object-oriented image analysis. *Studies in Computational (SCI), Springer-Verlag*, 126 :71–88, 2008.

E. B. Fowlkes et C. L. Mallows. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association, Taylor and Francis, Ltd.*, 78 :553 – 569, 1983.

A. Fred. Finding consistent clusters in data partitions. *3rd. Int. Workshop on Multiple Classifier Systems*, pages 309 – 318, 2001.

A. LN. Fred et A. K. Jain. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6) :835 – 850, 2005.

R. Ghaemi, M. N. Sulaiman, H. Ibrahim, et N. Mustapha. A survey : Clustering ensembles techniques. *World Academy of Science, Engineering and Technology*, 2009.

J. Ghosh et A. Acharya. Cluster ensembles. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 1(4) :305–315, 2011.

G. Giacinto, F. Roli, et L. Bruzzone. Combination of neural and statistical algorithms for supervised classification of remote sensing images. *Pattern Recognit. Lett.*, 21 :385–397, 2000.

A. Gordon et M. Vichi. Fuzzy partition models for fitting a set of partitions. *Psychometrika*, 66(2) :229 – 248, 2001.

D. Greene et P. Cunningham. Efficient ensemble methods for document clustering. *Technical report, Department of Computer Science, Trinity*, 2006.

M. Halkidi, Y. Batistakis, et M. Vazirgiannis. Clustering validity checking methods. *Part ii*, 2001.

G. Hamerly et C. Elkan. Alternatives to the k-means algorithm that find better clustering. *Proceedings of the 11th International Conference on Information and Knowledge Management*, page 600–607, 2002.

M. Han, X. Zhu, et W. Yao. Remote sensing image classification based on neural network ensemble algorithm. *Neurocomputing*, 78(1) :133–138, 2012.

J. Handl, J. Knowles, et D. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21 :3201 – 3212, 2005.

Y. Hong, S. Kwong, Y. Chang, et Q. Ren. Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm. *Pattern Recognition Society*, 41 :2742– 2756, 2008.

N. Iam-On, T. Boongeon, S. Garrett, et C. Price. A link-based cluster ensemble approach for categorical data clustering. *IEEE Transactions on Knowledge and Data Engineering*, 24(3) :413–425, 2012.

A.K. Jain et R.C. Dubes. Algorithms for clustering data. *Prentice-Hall, Inc., Upper Saddle River*, 1988.

G. Karthik et M. Sangeetha. Remote sensing satellite image classification using neural network. *Proceedings of the First International Conference on Computing, Communication and Control System, I3CAC 2021*, 2021.

L. Kaufman et P. Rousseeuw. Clustering by means of medoids. 1987.

S. Khedairia et M.T. Khadir. A multiple clustering combination approach based on iterative voting process. *Journal of King Saud University – Computer and Information Sciences*, 2019.

J. Kittler, M. Hatef, R.P.W. Duin, et J. Matas. On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(3) :226 – 239, 1998.

M. Kodinariya et R. Makwana. Review on determining number of cluster in k-means clustering. *Int. J. Adv. Res. Comput. Sci. Manage. Stud.*, 1(6) :90 – 95, 2013.

T. Kohonen. *Self-Organizing Maps*, volume 30. Springer, Germany, Berlin /Heidelberg, 1995.

K. Koonsanit, C. Jaruskulchai, et A. Eiumnoh. Parameter-free k-means clustering algorithm for satellite imagery application. *International Conference on Information Science and Applications*, 2012.

G. Kumar, P.P. Sarth, P. Ranjan, et S. Kumar. Satellite image clustering and optimization using k-means and pso. *IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*, 2016.

L.I. Kuncheva. Classifier ensembles : Facts, fiction, faults and future. *(slides, plenary talk)*, 2008.

L.I. Kuncheva. *Combining pattern classifiers, Methods and Algorithms*. John Wiley and Sons, Inc., Hoboken, New Jersey., 2014.

K. Labed, H. Fizazi, H. Mahi, et I.M. Galvan. A comparative study of classical clustering method and cuckoo search approach for satellite image clustering : Application to water body extraction. *Applied Artif. Intell.*, 32(1) :96 – 118, 2018.

T. Li, C. Ding, et M. Jordan. Solving consensus and semisupervised clustering problems using nonnegative matrix factorization. *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 577 – 582, 2007.

T. M. Lillesand et R. W. Kiefer. *Remote sensing and image interpretation*. Wiley, Chichester, 1994.

D. Lu et Q. Weng. A survey of image classification methods and techniques for improving classification performance. *Int. J. Remote Sens.*, 28 :823–870, 2007.

H. Luo, F. Jing, et X. Xie. Combining multiple clusterings using information theory based genetic algorithm. *Proceedings of the International Conference on Computational Intelligence and Security*, 1 :84 – 89, 2006.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, 1 :281–297, 1967.

H. Mahi, N. Farhi, et K. Labed. Remotely sensed data clustering using kharmonic means algorithm and cluster validity index. *Proceedings of the 5th IFIP TC5 International Conference on Computer Science and Its Applications, CIIA'2015, Saïda, Algeria*, 456 :105 – 116, 2015.

P.M. Mather. *Classifcation methods for remotely sensed data*. Taylor and Francis Group, LLC, 2009.

Y. Miao, H. Wang, et B. Zhang. Multiple classifier system for remote sensing images classification. *Intelligence Science and Big Data Engineering. Springer*, 11266 :491–501, 2018.

R.R. Navalgund, V. Jayaraman, et P.S. Roy. Remote sensing applications : An overview. *Current*, 93 :1747–1766, 2007.

O. Okun et G. Valentini. *Supervised and Unsupervised Ensemble Methods and their Applications*. Springer-Verlag Berlin Heidelberg, 2008.

M. K. Pakhira, S. Bandyopadhyay, et U. Maulik. Validity index for crisp and fuzzy clusters. *Pattern Recognition*, 37(3) :487–501, 2004.

A. Richards. *Remote Sensing Digital Image Analysis*. Springer, Berlin, 1993.

J. Ronald Eastman. *Guide to GIS and Image Processing*. Clark University, 2009.

P.J. Rousseeuw et A.M. Leroy. Robust regression and outlier detection. *John Wiley and Sons, Inc.*, 1987.

Floyd F. Sabins. *Remote sensing : principles and interpretation*. W. H. Freeman, San Francisco, 1978.

S. Sarumathi, N. Shanthi, S. Vidhya, et M. Sharmila. A comprehensive review on different mixed data clustering ensemble methods. *International Scholarly and Scientific Research and Innovation*, 8, 2014.

R.E. Schapire. The strength of weak learnability. *Mach. Learn.*, 5(2) :197 – 227, 1990.

Gert A. Schultz et Edwin T. Engman. *Remote Sensing in Hydrology and Water Management*. Springer, Berlin, Heidelberg, 2000.

P.C. Smits. Multiple classifier systems for supervised remote sensing image classification based on dynamic classifier selection. *IEEE Trans. Geosci. Remote Sens.*, 40 :801–813, 2002.

B.M. Steele. Combining multiple classifiers : An application using spatial and remotely sensed information for land cover type mapping. *Remote Sens. Environ.*, 74 :545–556, 2000.

M. Steinbach, G. Karypis, et V. Kumar. A comparison of document clustering techniques. *Workshop on Text Mining, KDD*, 2000.

A. Strehl et J. Ghosh. Cluster ensembles—a knowledge reuse framework for multiple partitions. *The Journal of Machine Learning Research*, 3 :583–617, 2002.

H. Su et P. Du. Multiple classifier ensembles with band clustering for hyperspectral image classification. *European J. Remote Sensing*, 47(I) :217–227, 2014.

J. Sublime, N. Grozavu, G. Cabanes, Y. Bennani, et A. Cornuejols. From horizontal to vertical collaborative clustering using generative topographic maps. *Int. J. Hybrid Intell. Syst.*, 12(4) :245–256, 2016.

H. J. Sun, S. R. Wang, et Q. S. Jiang. Fcm-based model selection algorithms for determining the number of clusters. *Pattern Recognition*, 37 :2027–2037, 2004.

P. Tan, M. Steinbach, et V. Kumar. *Introduction to data mining*. Pearson Addison Wesley, Boston, 2006.

P.N. Tan, M. Steinbach, et V. Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc, Boston, MA, USA, 2005.

V.V. Tatarnikov, I.A. Pestunov, et V.B. Berikov. Centroid averaging algorithm for a clustering ensemble. *Comput. Optics*, 41(5) :712–718, 2017.

T. Thirupathi et D. Nagasudha. Neural network based terrain classification using remote sensing. *International Journal of Advanced Science and Technology*, 29 :2050 – 2064, 2020.

A. Topchy, A.K. Jain, et W. Punch. Combining multiple weak clusterings. *Proceeding of the Third IEEE International Conference on Data Mining*, 2003.

A. Topchy, A.K. Jain, et W. Punch. Clustering ensembles : Models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12) :1866–1881, 2005.

A. *Topchy*[a], A.K. Jain, et W. Punch. A mixture model of clustering ensembles. *In Proceedings of the SIAM International Conference of Data Mining. Citeseer*, 2004.

A. *Topchy*[b], B. Minaei Bidgoli, A. K. Jain, et W. Punch. Adaptive clustering ensembles. *Proceeding International Conference on Pattern Recognition (ICPR)*, pages 272 – 275, 2004.

K. Tumer et A. Agogino. Ensemble clustering with voting active clusters. *Patt. Recogn. Lett.*, 29 :1947 – 1953, 2008.

B. Usman. Satellite imagery land cover classification using k-means clustering algorithm. computer vision for environmental information extraction. *Comput. Sci. Eng.*, 63 :18671–18675, 2013.

S. Vega-Pons, J. Correa-Morris, et J. Ruiz-Shulcloper. Weighted partition consensus via kernels. *Pattern Recognition*, 43(8) :2712 – 2724, 2010.

S. Vega-Pons et J. Ruiz-Shulcloper. A survey of clustering ensemble algorithms. *Int. J. Pattern Recogn. Artif. Intell.*, 25(3) :337–372, 2011.

B. Waske, S. van der Linden, J.A Benediktsson, A. Rabe, et P. Hostert. Sensitivity of support vector machines to random feature selection in classification of hyperspectral data. *IEEE Trans. Geosci. Remote Sens.*, 48 :2880–2889, 2010.

C. Wemmert. *Classification hybride distribuée par collaboration de méthodes non supervisée*. PhD thesis, Université de Strasbourg, 2000.

W.L. Winston et J.B. Goldberg. *Operations Research : Applications and Algorithms*. 4rd ed. Duxbury Press, Belmont, California, 2004.

D.H. Wolpert. Stacked generalization. *Neural networks : the official journal of the International Neural Network Society*, 5(2) :241 – 259, 1992.

Y. Yang. *Unsupervised ensemble learning and its application to temporal data clustering*. PhD thesis, University of Manchester, 2011.

Y. Yang. *Temporal Data Mining via Unsupervised Ensemble Learning*. 1st ed. Elsevier, Amsterdam, 2017.

H. *Yoon*[a], S. Ahn, S. Lee, S. Cho, et J. H. Kim. Heterogeneous clustering ensemble method for combining different cluster results. *Data Mining for Biomedical Applications, Springer*, pages 82– 92, 2006.

H. *Yoon*[b], S. Lee, S. Cho, et J. H. Kim. A novel framework for discovering robust cluster results. *DS 2006, LNAI, Springer-Verlag Berlin Heidelberg*, 4265 : 373 – 377, 2006.

D. Yu, Q. Xu, H. Guo, C. Zhao, Y. Lin, et Li. An efficient and lightweight convolutional neural network for remote sensing image scene classification. *sensors*, 20, 2020.

B. Zhang. Generalized k-harmonic means. boosting in unsupervised learning. *Technical Reports HPL-2000-137, Hewlett - Packard Labs*, 2000.

Y. Zhang. Ten years of technology advancement in remote sensing and the research in the crc-agip lab in gce. *Geomatica*, 64 :173–189, 2010.

Q. Zhao. *Cluster validity in clustering methods*. Publications of the University of Eastern Finland. Dissertations in Forestry and Natural Sciences, 77, 2012.

Q. Zhao et P. Fränti. Wb-index : a sum-of-squares based index for cluster validity. *Knowledge and Data Engineering*, 92 :77 – 89, 2014.

Y. Zhao et G. Karypis. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10(2) :141 – 168, 2005.

J. Zheng, Z. Cui, A. Liu, et Y. Jia. A k-means remote sensing image classification method based on ada boost. *The 4th International Conference on Natural Computation (ICNC '08)*, pages 27–32, 2008.

Z. Zhou et W. Tang. Cluster ensemble. *Knowledge-Based Systems*, 19(1) :77 – 83, 2006.

Résumé

En télédétection, le clustering, également appelée classification non supervisée, est une tâche importante qui vise à partitionner une image donnée dans un espace multispectral en un certain nombre de classes spectrales (groupes), lorsque l'information in situ n'est pas disponible. Parmi les nombreux algorithmes de clustering existants, les plus utilises sont le K-means, l'ISODATA, le FCM (Fuzzy C-Means), le SOM (Self Organizing Map) et plus récemment le K-Harmonic Means. Cependant, avec l'augmentation de la quantité de données détectées à distance et leur hétérogénéité, il devient difficile d'obtenir des résultats de clustering pertinents en utilisant un seul algorithme. De plus, chaque algorithme précité nécessite un certain nombre de paramètres et le plus important d'entre eux est le nombre de clusters, que l'utilisateur doit définir a priori.

Pour faire face à ces lacunes, les systèmes de classifications multiples (MCS), également connus sous le nom d'ensemble de clustering, est le consensus de différents algorithmes de clustering qui peut fournir la meilleure partition avec une grande précision et, par conséquent, surmonter les limites des approches traditionnelles basées sur des classificateurs uniques. Le MCS comprend deux étapes : la génération de partitions et la combinaison de partitions.

Dans cette thèse, nous étudions les avantages et les potentiels de cette technique dans le domaine de l'occupation du sol en utilisant différents types de données : Données synthétiques, données composites et données de télédétection. La première étape du MCS est assurée par quatre algorithmes de clustering, à savoir l'algorithme k-means, l'algorithme k-harmonic means (KHM), l'algorithme Bisecting K-means (BKM) et l'algorithme Self Organizing Map (SOM). Le meilleur clustering qui fait office de référence est obtenu selon l'indice WB. Les méthodes de ré-étiquetage et de vote sont utilisées dans la deuxième étape. Les résultats expérimentaux obtenus par le MCS surpassent légèrement les résultats du clustering individuel.

Mots clés : Clustering, K-means, k-harmonic means, Bisecting K-means, Self Organizing Map, indices de validité des clusters, données de télédétection.

Abstract

In remote sensing, clustering, also called unsupervised classification, is an important task that aims to partition a given image in a multispectral space into a number of spectral classes (clusters), when in situ information is not available. Among the many existing clustering algorithms, the most commonly used are K-means, ISODATA, FCM (Fuzzy C-Means), SOM (Self Organizing Map) and more recently K-Harmonic Means. However, with the increase in the amount of remotely sensed data and its heterogeneity, it becomes difficult to obtain relevant clustering results using a single clustering algorithm. Moreover, each algorithm mentioned above requires a number of parameters and the most important of them is the number of clusters, which the user has to define a priori.

To cope with these shortcomings, the Multiple Classifier System (MCS) is also known as ensemble clustering, is the consensus of different clustering algorithms can provide the best partition with high accuracy and consequently overcome limitations of traditional approaches based on single classifiers. The MCS involves two stages: the partitions generation and the partitions combination.

In this thesis, we investigate the potential advantages of this technique in the unsupervised land cover classification by using various kinds of data: Synthetic data, composite data and remotely sensed data. The first stage of the MCS is assumed by four clustering algorithms, the well-known k-means algorithm, the k-harmonic means algorithm (KHM), Bisecting K-means (BKM) and the self-organizing map (SOM). The best clustering is obtained according to WB index. The relabeling and the voting methods are used in the second stage. Experimental results obtained by the MCS outperform the results of the individual clustering.

Keywords: Clustering, K-means, k-harmonic means, Bisecting K-means, self-organizing map, cluster validity indices, remotely sensed data.

ملخص

في الاستشعار عن بعد ، يعتبر التجميع (clustering) ، الذي يُطلق عليه أيضًا التصنيف غير الخاضع للإشراف (classification non supervisée)، مهمة أساسية تهدف إلى تقسيم صورة معينة في مساحة متعددة الأطياف إلى عدد من الفئات (المجموعات) الطيفية ، عندما لا تتوفر المعلومات في الموقع. من بين العديد من خوارزميات التجميع (clustering) الحالية الأكثر استخدامًا هي : K-mean ، ISODATA، FCM (Fuzzy C-Means)، SOM (Self Organizing Map) ومؤخراً K-Harmonic Means. ومع ذلك ، مع الزيادة في كمية البيانات المكتشفة عن بعد وعدم تجانسها ، يصبح من الصعب الحصول على نتائج التجميع (clustering) ذات الصلة باستخدام خوارزمية واحدة. بالإضافة إلى ذلك ، تتطلب كل خوارزمية سالفة الذكر عددًا من العوامل وأهمها عدد المجموعات التي يجب على المستخدم تحديدها مسبقًا.

للتعامل مع هذه النقائص ، فإن أنظمة التصنيف المتعددة (MCS) ، والمعروفة أيضًا باسم مجموعة التجميع (ensemble de clustering) ، هي توافق خوارزميات التجميع (clustering) المختلفة التي يمكن أن توفر أفضل تجزئة بدقة عالية وبالتالي التغلب على قيود الأساليب التقليدية القائمة على المصنفات الفردية. يتكون MCS من خطوتين: إنشاء التجزئات والجمع بين التجزئات .

في هذه الأطروحة ، ندرس مزايا وإمكانيات هذه التقنية في مجال الغطاء الأرضي باستخدام أنواع مختلفة من البيانات: البيانات التركيبية والبيانات المركبة وبيانات الاستشعار عن بعد. في المرحلة الأولى من MCS يتم الإعتماد على أربع خوارزميات تجميع (clustering)، وهي خوارزمية k-means ، خوارزمية k-harmonic means (KHM)، خوارزمية Bisecting K-mean (BKM) وخوارزمية خريطة التنظيم الذاتي Self Organizing Map (SOM). إن أفضل تجميع مرجعي تم الحصول عليه وفقا لمؤشر WB. في الخطوة الثانية يتم استخدام طريقتي إعادة وضع العلامات (ré-étiquetage) والتصويت. النتائج التجريبية التي تم الحصول عليها بواسطة MCS تجاوزت بشكل طفيف نتائج التجميع الفردي (clustering individuel) .

الكلمات المفتاحية : التجميع (clustering) ، K-means، k-harmonic means ، K- Bisecting means، خريطة التنظيم الذاتي (Self Organizing Map) ، مؤشرات صلاحية المجموعات (indices de validité des clusters)، بيانات الاستشعار عن بعد (données de télédétection).