République Algérienne Démocratique et Populaire
Ministère de L'enseignement Supérieur et de la Recherche Scientifique
Université 8 Mai 1945 - Guelma
Faculté des Sciences et de la Technologie
Département Electronique et Télécommunications

Mémoire de fin d'étude
Pour l'obtention du diplôme de Master Académique

Domaine : **Sciences et Techniques**
Filière : **Electronique**
Spécialité : **Systèmes Electroniques**

Validation des méthodes hiérarchique

Présenté par :
SAHRAOUI Abdelhakim
TAHRI Salem

Sous la direction de : Dr BOUDOUDA Houria

JUIN 2011


جامعة 08 ماي 45 قالمة
العلوم والتكنولوجيا
11/2701

Remerciements

Premièrement, nous tentons à remercier DIEU, le tout puisant qui a éclairé notre chemin, et la patience qu'il nous a donné pour réaliser ce travail.

En guise de reconnaissance et gratitude, nos sincères remerciements et notre profonde reconnaissance sont adressés à toutes les personnes qui ont œuvré de prés ou de loin à l'accomplissement de ce mémoire, notamment :

Au Dr. BOUDOUDA HOURIA qui nous a aidés du début à la fin de l'élaboration de ce mémoire de fin cycle.

Au département de Télécommunication et Electronique par la voix de son directeur qui a su mettre à notre disposition les structures nécessaires pour la réalisation de ce mémoire.

A l'université de Guelma pour leur accueil chaleureux à notre égard ainsi que les dispositions mises pour nos études.

Toutes nos familles pour tout le soutien qu'elles nous ont accordé tout au long de notre parcours et les moyens nécessaires mis pour nos éducations et nos formations.

DÉDICACES

C'est avec un immense honneur et une grande modestie que je dédie ce modeste travail à :

Mes parents :

Pour toute la tendresse, amour et affection qui ont été pour moi une lumière et un appui d'une valeur inestimable. Ici, je prêt à mes parents le témoignage de mes sentiments les plus distinguer et s'il y a quelqu'un au monde envers qui je dois beaucoup, ça serait mes parents et quoique je fasse jamais je ne pourrai revaloir ce que vous m'avez donné avec cœur et âme.

À notre encadreur Dr. BOUDOUDA HOURIA

A mon ami ce Travail Salem Tahri

À tous mes amis de classe surtout **Mohamed Rahman**

À mes frères : ISHAK et FAKHRI.
À mes sœurs : LAMIA et NOUR ELHOUDA et SALIMA.

À mes oncles : Ahmed, abdelbasset, Hamdi, Bouka

Et les feus : mon encles Alsadék, grand-père Massoud, grand-mère ouarida

À mes cousins ABDELATIF et MASSOUD et AYMEN et KHALED.

Et Tout le reste de ma Famille

À mes amis : Salim, Belgacem, Habib, sabri ,Hamid ,Azzedine ,Bilel ,Majid ,Tarek ,Sami ,Faouzi ,abdelhak ,Mohamed ,Kamel , Ouehab ,dénna.
Et tout qui j'oubliai.

Abdelhakim

DÉDICACES

C'est avec un immense honneur et une grande modestie que je dédie ce modeste travail à :

Mes parents :

Pour toute la tendresse, amour et affection qui ont été pour moi une lumière et un appui d'une valeur inestimable. Ici, je prêt à mes parents le témoignage de mes sentiments les plus distinguer et s'il y a quelqu'un au monde envers qui je dois beaucoup, ça serait mes parents et quoique je fasse jamais je ne pourrai revaloir ce que vous m'avez donné avec cœur et âme.

À notre encadreur: Dr BOUDOUDA HOURIA.

A mon ami ce Travail: Sahraoui Abdelhakim.

À tous mes amis de classe : (Badri, Nabil, Med, Brahim, Rouf, Amine, anis et Haroun).

À mes frères: Nadji et Abdelmalek.
À mes sœurs: Malika et Fatiha.

Enfants de mon frère: Amer, Seyfe.

À mes oncles: Med, Medtaher et feu Joseph.

Et les feus: Mon grand-père Ahmed, Mon grand-père Younes, Ma grand-mère Hadda et Ma grand-mère Mabrouka.

À mes cousins: Med, Mahmoud, Abderrahmane, Kamal, Islame…

Et Tout le reste de ma Famille: Tahri, salhi et bennadji.

À mes amis: Mounir, Hakim, Rachid, Riadh, Mansour, Rabie, Naserddine, Abdelmoumène, Hafed, feu Ali Farrouge, Younes, Nadjib, Adel, Ahmed, Abdeldjalil, Ouanis, Nauare, Maftah, Med S, Hocine, Ridha, Med T, Othmane, Saber, Khaled, Omar, Salah, Sami, Abderrahmane, Azddine, Djalale.

Salem

Table des matières

Résumé :

Le présent document décrit le travail réalisé pour l'obtention du master en Electronique à l'université de Guelma. Le travail consiste à faire une présentation de des méthodes de classification hiérarchique ascendantes : SLINK, ALINK, GALINK, CALINK, Ward. L'objectif est la validation de ces méthodes en utilisant : La corrélation cophnetic et l'inconsistance. Des tests ont été réalisés sur deux bases données pour valider les résultats de la classification. La simulation montre la robustesse du ALINK et GALINK appliqués sur une distance Euclidienne et concordance des deux indices de validité.

Chapitre 01

Reconnaissance des formes

L'idée de construire des machines capables de simuler des êtres humains afin de les aider dans certaines tâches, voire de les remplacer, était antérieure aux ordinateurs. Leur apparition a permis d'étendre le spectre des tâches à simuler en ajoutant celles dont l'exécution relève de facultés mentales comme la perception et le raisonnement.

Le problème que cherche à résoudre la reconnaissance des formes est d'associer une étiquette à une donnée. A cet effet, plusieurs méthodes ont été développées pour bien comprendre la structure des données en assignant automatiquement des données à des classes différentes (reconnaissance).

Parmi ces méthodes, les méthodes hiérarchiques produisent une hiérarchie complète qui est une séquence imbriquée de partitions de données d'entrée. Elles peuvent être soit d'agglomérations (ascendantes) ou de division (descendantes). Les méthodes ascendantes génèrent une séquence de partitions imbriquées en partant d'un regroupement trivial dans lequel chaque élément se trouve dans une classe unique et en terminant par le regroupement trivial où tous les éléments sont dans le même cluster. Une méthode de division, comme son nom l'indique, effectue une procédure de division partant d'un cluster regroupant tous les objets jusqu'à ce qu'un critère d'arrêt soit atteint (généralement jusqu'à l'obtention d'une partition de clusters représentés par des singletons).

Considérant une méthode hiérarchique ascendante, après l'organisation de touts les objets dans un arbre hiérarchique, on passe ensuite à l'étape de validation. Pour cela on doit vérifier d'une part, si les distances entre les classes dans l'arbre reflètent ou non les distances d'origine entre les objets. D'autre part, on doit chercher les divisions naturelles qui existent entre les liens. Deux mesures sont disponibles :

1. *Coefficient de corrélation* cophenetic
2. *Coefficient d'inconsistance*

Dans l'arbre hiérarchique généré et à un certain niveau, deux objets quelconques dans la base de données initiale sont par la suite relies. La hauteur du lien représente la distance entre les deux classes qui contiennent ces deux objets. Cette hauteur est connue comme distance 'cophenetic' entre les deux objets.

I.1 Introduction

La reconnaissance de formes (RDF) [Duda, 73] est l'une des rares disciplines qui n'a pas besoin de définition formelle dans le cas général. Les mots reconnaissance et forme sont dans le vocabulaire courant de tous les jours.

Pour l'ingénieur, il s'agit bien étendu de l'étude des systèmes automatiques ou semi-automatiques capables de reconnaître les formes qu'on leur présente.
Sans nul doute, l'homme est le plus parfait des systèmes de reconnaissance de formes que nous pouvons accomplir sur des formes à grande variabilité.

Les progrès scientifiques et technique nous permettant aujourd'hui d'essayer d'imiter certaines de ces facultés à l'aide des machines. Malgré tout, beaucoup de cas particulier ou les travaux de reconnaissance simple effectués par l'homme peuvent être confiés à une machine.

Les motivations de cette substitution se situent sur trois plans principaux qui sont ceux de l'efficacité, de social et de l'économie.

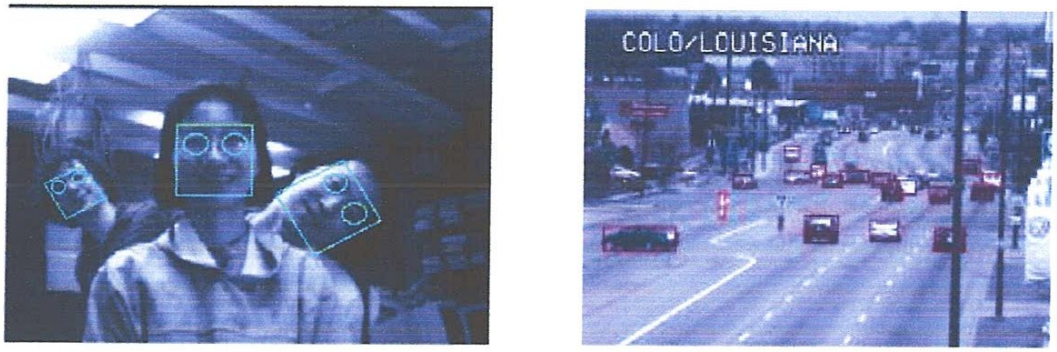


Figure I.1 : Exemples de la reconnaissance.

I.2 Domaines d'application de la RDF

I.2.1 Domaine industriel

- La robotique généralement utilisé dans l'industrie ;
- L'assemblage pour servir la reconnaissance de pièces et leur montage ;
- Le contrôle de qualité des produits et dans le processus de production ;

- La fabrication des véhicules autonomes dans les industries automobiles et divers ainsi que les pièces de rechange etc....

I.2.2 Les systèmes de télédétections

- La météo (tempête, ouragan...) ;
- L'identification et le suivi des cultures, des forêts, des réserves d'eau ;
- La cartographie ;
- L'analyse des ressources terrestres.

I.2.3 La médecine

- L'analyse de l'ECG ou EEG pour fin de diagnostic ;
- L'analyse d'images médicales ;
- Echographie, IR, Microscope (histologie,... etc.) pour détecter des tumeurs, cancéreuses ou autres maladies.

I.2.4 Application militaire

- Guidage de missile (reconnaissance d'une cible et du terrain) ;
- Reconnaissance aérienne (espionnage) ;

I.2.5 La bureautique

- La reconnaissance de texte par ordinateur (OCR) ;
- L'analyse de document ;
- La reconnaissance de la parole.

I.2.6 La sécurité

- L'identification des empreintes digitales (iris, main) ;
- La reconnaissance de visage ;
- L'authentification de la parole ;
- L'identification de signature.

I.3 Le schéma de la reconnaissance de formes

L'objectif de la reconnaissance de formes est de classifier des entités en catégories à partir d'observations et effectuées sur celles-ci.Ce dispositif se décompose généralement en 5 étapes (voir figureI.2) :

- Acquisition des données;
- Génération de caractéristiques;
- Extraction/Sélection des caractéristiques;
- Classification ;
- Évaluation du système.

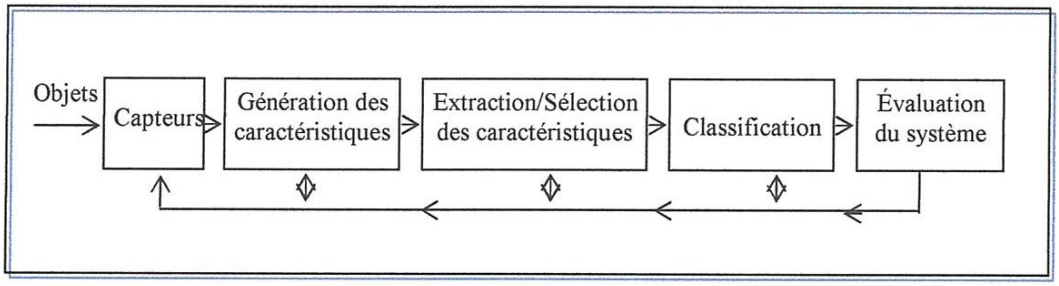


Figure I.2: Processus de la reconnaissance de formes.

I.4 L'acquisition des données

De l'acquisition des données à la classification, la qualité du système dépend fortement des étapes précédentes. Pour illustrer le déroulement standard de ce processus, prenons l'exemple d'un système d'aide à la conduite automobile dont l'objectif est de prévenir le conducteur d'un éventuel danger par l'intermédiaire de divers capteurs.

Dans une application, les données acquises peuvent être de type et de nature différente et dépendent de ce que l'on recherche. Dans le cas de notre exemple, on peut avoir :

- Des informations sur l'état du véhicule telles que la position, la vitesse, l'accélération, pression de freinage, etc. ... (Capteurs proprioceptifs) ;

- Des informations sur l'environnement du véhicule avec le GPS, la télémétrie, des caméras, etc.;
- Des informations propres au véhicule comme le modèle, la couleur, la plaque minéralogique, etc. ... ;
- Des informations sur le conducteur (hypovigilance, âge, taille, sexe).

I.5 Les différentes méthodes de la RDF

Les différentes méthodes de la RDF souvent sont regroupées en grandes classes identifiées par : statistique, syntaxique, structurelle, hybride (une combinaison des autres).

I.5.1 Reconnaissance de formes structurelle

C'est une approche qui est basée sur l'extraction de primitives en prenant compte de l'information structurelle. Elle cherche à structurer l'information en décrivant l'organisation topologie (la structure) de la forme à partir de ses composantes les plus élémentaires. Cette approche nécessite une mesure de la similarité entre deux représentations structurelles. On distingue plusieurs techniques telles que les structures des graphes et les structures syntaxiques.

I.5.2 Reconnaissance des formes statistique

Cette approche consiste à déterminer des caractéristiques extraites d'une forme pour les caractériser d'une manière statistique [Andrw, 99]. Elle a besoin d'un nombre élevé d'exemples afin de réaliser un apprentissage correct des lois de probabilité des différentes classes. Autrement dit, cette approche bénéficie des méthodes d'apprentissage automatique qui s'appuient sur des bases théoriques connues telles que la théorie de la décision bayésienne, les méthodes de classification non supervisées et l'analyse en composantes principales. Les deux principales familles de méthodes utilisées sont les méthodes paramétriques et les méthodes non paramétriques.

a. Méthode paramétrique

Les méthodes paramétriques opèrent sous l'hypothèse que les classes étudiées suivent une distribution de probabiliste d'une certaine forme connue à priori. La prise de décision consiste à déterminer la classe pour laquelle la forme inconnue présente la probabilité d'appartenance maximale. Elles exigent des bases d'apprentissage assez importantes pour une estimation correcte des paramètres de la distribution supposée.

L'approche statistique englobe : la règle de Bayes, la distance de Mahalanobis , les méthodes neuronales et les chaînes de Markov .

b. Méthode non paramétrique

Dans le cas des méthodes non paramétriques, les lois de probabilité sont inconnues pour une des classes. Le problème revient à établir des frontières de décision entre les classes. Les techniques les plus utilisées en reconnaissance de formes sont : la méthode du plus proche voisin, la méthode de Parzen et la méthode d'appariement de graphes. Pour de plus amples informations, Gaillat décrit un ensemble de méthodes statistiques en reconnaissance de formes.

Malgré leur nature différente, les approches statistiques et structurelles peuvent être combinées aux mêmes domaines d'application. Le choix d'une approche peut être lié à des contraintes matérielles telle que la taille de la base d'apprentissage disponible, le temps de calcul requis et la taille mémoire nécessaire. L'utilisation conjointe des deux approches peut être une solution optimale pour le problème de reconnaissance de l'écriture.

I.6 Conclusion

Un système de reconnaissance de formes peut comporter comme une phase d'apprentissage qui va consister à « apprendre » à reconnaître des formes sur la base d'échantillons. Lorsque cette phase d'apprentissage sera achevée le système sera alors prêt à fonctionner pour reconnaître des formes inconnues qui lui seront soumises, mais un système de reconnaissance de formes peut être aussi un système qui trie (fait des «paquets» homogènes suivant certains critères) un ensemble de formes inconnues. Il n y a alors pas d'apprentissage à proprement parler.

Chapitre 02

La classification hiérarchique

II.1 Introduction

Comme tout principe de classification non supervisée la classification hiérarchique détermine une structuration des données en regroupant celles qui possèdent des propriétés similaires. Cependant elle ne s'arête pas à cette structuration « horizontale » en classes. Elle cherche aussi à établir un lien hiérarchique entre les regroupements.

Il existe deux méthodes principales :

- Les méthodes de classification hiérarchique ascendante (CHA)
- Les méthodes de classification hiérarchique descendantes(CHD)

II.2 Notions élémentaires sur la classification

Dans cette partie, les principes généraux liés à classification sont décrits de façon formelle pour illustrer les concepts et mettre en avant un certain nombre d'écueils.

II.2.1 Espaces des données et des catégories

Appelons K la fonction de classification idéale que l'on veut réaliser, du moins approximativement, par un système automatique. Pour un objet donnéx, la valeur K(X) est l'indice(l'étiquette,) ou la catégorie à laquelle doit être affecté.

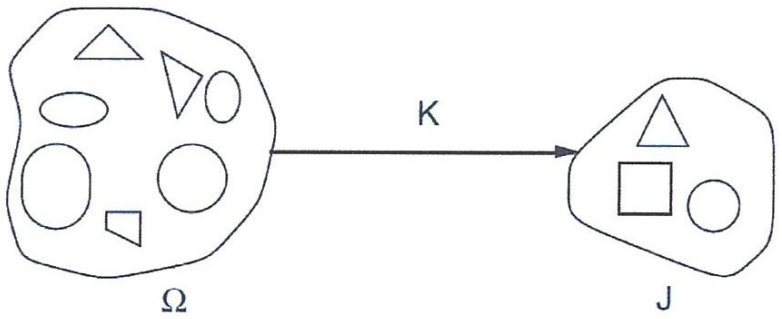


Figure II.1: Classification de formes en trois catégories

Ω: Espace des données ou des formes, population des objets analysés.
J : espace des indices, codes ou identificateurs qui désignent les catégories servant à classer les objets donnés.
K = fonction de classification groupant les objets donnés en catégories disjointes.

II.2.2 Représentation des données

Les données d'un problème de classification peuvent être vues comme un tableau de *N* lignes (*N* étant le nombre d'objets) et *P* colonnes (*P* étant le nombre d'attributs : dimension de l'espace des données). Un objet (ou individu) sera donc un vecteur noté *X* ayant *M* composantes (les valeurs des attributs de l'objet). On notera $\vec{X}_i$ le i^e objet de la base de données et x_{ik} la valeur de son k^e attribut. L'ensemble des données peut alors être défini par :

$$X = \left\{\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_N\right\}$$

II .2 .3 Mesures de Similarités

L'une des problématiques centrales de la classification est de définir la notion de similarité « rapprochement» entre les données. Il existe trois concepts de similarité en classification : la similarité entre objets à maximiser pour deux objets appartenant au même classe, et à minimiser pour deux objets appartenant à des classes différents; la similarité entre un objet et un classe à maximiser si l'objet est associé au classe pour une bonne *cohésion interne* du classe; et la similarité entre classes à minimiser pour une bonne *isolation externe* des classes. La notion de rapprochement est centrale dans les méthodes de classification. Il est donc important de bien comprendre sa raison d'être, ses variantes et les impacts du choix d'une mesure particulière sur les classes.

II.2.3.1 Similarité ente objets

Typiquement, la similarité entre objets est évaluée par une fonction de distance définie entre paire d'objets. En outre, Il est évident que le type des données influence la manière de mesurer le rapprochement potentiel de deux objets. Dans notre cas, on considère des objets dont tous les attributs prennent des valeurs numériques, peu importe que celles-ci soient continues, discrètes ou par intervalles. Les mesures de distance les plus courantes sont les suivantes :

a. Distance de Manhattan ou city-bloc

$$D(X_i, X_j) = \sum_{k=1}^{M} \left| x_{ik} - x_{jk} \right|$$

A partir de cette définition, on peut s'interroger sur la forme des classes qu'il est possible de détecter dans l'espace des données. Pour cela, considérons le cas simple d'un ensemble d'objets bidimensionnels (c'est-à-dire décrits par deux attributs). Le lieu des points situés à égale distance γ d'un centre (c,d) dans le plan vérifie alors l'équation

$$|x-c|+|y-d|=\gamma$$

On montre assez facilement que ce lieu correspond à la représentation de la figure 2, la diagonale du carré étant égale à 2.

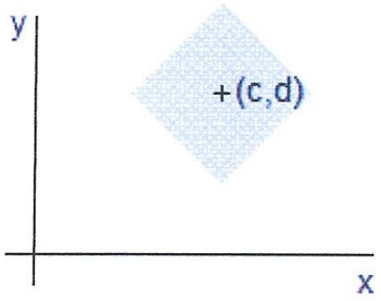


Figure II.2: Visualisation d'une classe par distance de Manhattan.

b. Distance Euclidienne

$$D(X_i, X_j)=\sqrt{\sum_{k=1}^{M}\left|x_{ik}-x_{jk}\right|^2}$$

La distance euclidienne correspond à la distance la plus couramment utilisée. Afin de s'intéresser à la forme des classes détectables par l'utilisation d'une telle distance, on reprend le même exemple que précédemment. On trouve l'équation du lieu des points :

$$\sqrt{(x-c)^2+(y-d)^2}=\gamma$$

Qui correspond à l'équation d'un cercle de centre (c,d) et de rayon γ dans le plan, représenté à la figure 3.

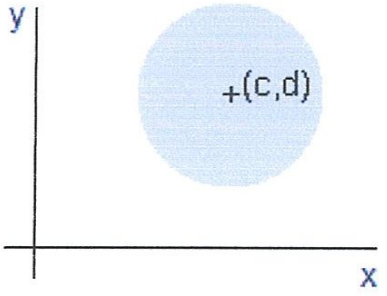


Figure II.3: Visualisation d'une classe par distance euclidienne

On voit donc que cette distance nous permettra de détecter des classes rondes dans un espace bidimensionnel, plus généralement des classes M-sphériques dans un espace à M dimensions.

c. Distance de Minkowski :

$$D(X_i, X_j) = \left(\sum_{k=1}^{M} \left| x_{ik} - x_{jk} \right|^R \right)^{1/R}$$

Cette distance généralise les deux précédentes. La forme des classes détectables pour différentes valeurs de R est reprise à la figure 4, toujours pour un exemple bidimensionnel.

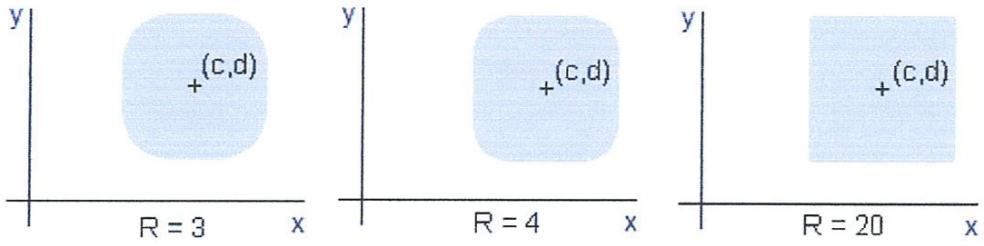


Figure II.4: Visualisation de classes par distance R

On constate que plus la valeur de R augmente, plus les classes détectables se rapprochent de carrés parfaits. De manière générale, la distance de Minkowski (et donc les distances de Manhattan, euclidienne, ...) fournissent de bons résultats lorsque les classes à détecter sont compacts et bien isolés.

d. Distance de Chebychev :

La distance de Chebychev est la limite de Minkowski pour R tendant vers l'infini.

$$D(X_i, X_j) = \lim_{R \to \infty} \left(\sum_{k=1}^{M} \left| x_{ik} - x_{jk} \right|^R \right)^{1/R}$$

Les classes détectables correspondent donc à des carrés parfaits.

e. Cosinus (Ochini coefficient)

$$\cos(X_i, X_j) = \frac{\sum_{k=1}^{M} x_{ik} \cdot x_{jk}}{\sqrt{\sum_{k=1}^{M} x_{ik}^2} \cdot \sqrt{\sum_{k=1}^{M} x_{jk}^2}} = \frac{\vec{X}_i \cdot \vec{X}_j}{\left\|\vec{X}_i\right\| \left\|\vec{X}_j\right\|}$$

Dans ce cas, le cosinus donne une mesure de similarité entre les deux objets. Une mesure de distance peut être simplement obtenue en prenant l'arc cosinus de cet indice de similarité.

II.2.3.2 Similarité de deux classes

Contrairement à la similarité entre objets, la similarité de deux classes, ne nécessite pas de calcule complexe et ne fait intervenir que des concepts physiques. Supposons avoir choisi une manière de mesurer le rapprochement de deux objets (peu importe laquelle). Le problème de déterminer celui de deux classes revient alors simplement à déterminer quels objets prendre dans chacun des classes pour définir la mesure. Les possibilités les plus courantes sont reprises ci-dessous.

a. Lien simple (SLINK)

Cette approche est encore nommée « *nearest neighbor approach* », ce qui traduit peut-être mieux son principe. Il s'agit donc de définir la distance entre deux classes comme étant la plus petite distance parmi celles entre toutes les paires d'objets entre les deux classes. Mathématiquement, la distance entre le classe C_p et le classe C_q est la plus petite distance entre un élément de C_p et un élément de C_q :

$$D(C_p, C_q) = \min\{dist(X_i, X_j), X_i \in C_p, X_j \in C_q\}$$

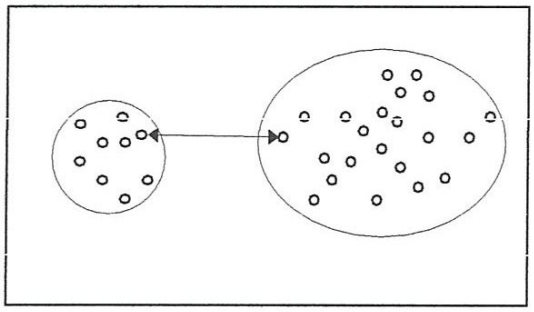


Figure II. 5: Distance entre deux classes par SLINK

b. Lien complet (CLINK)

Cette approche est encore nommée « *farthest neighbor approach* », ce qui traduit peut-être mieux son principe. Il s'agit donc de définir la distance entre deux classes comme étant la plus grande distance parmi celles entre toutes les paires d'objets entre les deux classes. Mathématiquement, la distance entre le classe C_P et le classe C_q est la plus grande distance entre un élément de C_P et un élément de C_q :

$$D(C_p, C_q) = \max\{dist(X_i, X_j), X_i \in C_p, X_j \in C_q\}$$

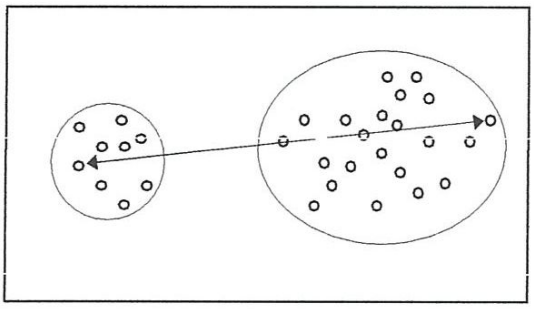


Figure II.6: Distance entre deux classes par CLINK

c. Lien moyen (ALINK)

Cette approche définit la distance entre deux classes en faisant intervenir tous les objets présents dans ces classes. Mathématiquement, la distance entre le classe C_P et le classe C_q est la moyenne des distances entre un élément de C_P et un élément de C_q :

$$D(C_p, C_q) = \frac{\sum_i \sum_j \{dist(X_i, X_j)\}, X_i \in C_p, X_j \in C_q\}}{cardinal(C_p).cardinal(C_p)}$$

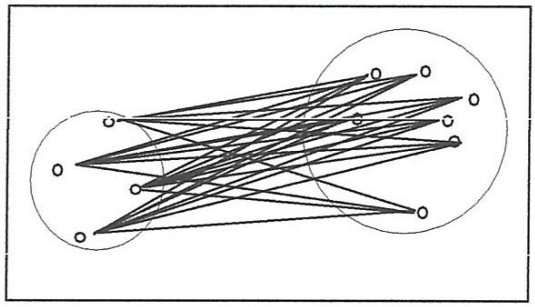


Figure II.7: Distance entre deux classes par ALINK

d. Lien moyen de groupe (GALINK)

Cette approche est encore nommée «méthode des centroïdes», ce qui traduit peut-être mieux son principe. Il s'agit donc de définir la distance entre deux classes comme étant la distance entre les centroïdes de ces classes. Mathématiquement, si G_P est le centroïde du classe C_p et si G_p est le centroïde du classe C_p alors la distance entre le classe C_p et le classe C_p est la distance entre leurs centroïdes :

$$D(C_p, C_p) = dist(GP, Gq)$$

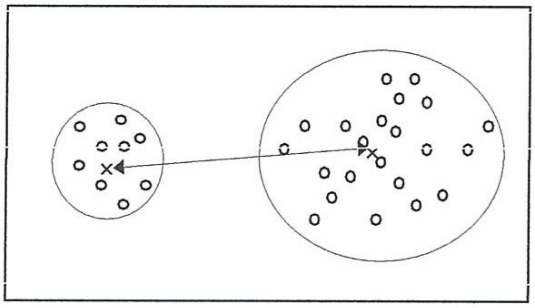


Figure II.8: Distance entre deux classes par GALINK

Remarquons qu'il n'est pas toujours nécessaire d'avoir recours à des mesures de rapprochements de classe. Ceci est surtout utile dans les algorithmes de classification hiérarchique ascendante.

II.2.3.3 Rapprochement d'un objet et d'une classe

Si l'on suppose avoir défini une mesure du rapprochement de deux objets, la définition de celui d'un objet et d'une classe est presque immédiate puisqu'il suffit de choisir un point représentatif de la classe. On se ramène donc à la mesure du rapprochement de deux objets : l'objet en question et le représentatif du classe, pouvant être son centroïde, l'objet le plus proche, l'objet le plus loin, ...

II.3 Procédure de classification

La procédure de classification consiste l'ensemble des étapes devant être réalisées depuis la mise à disposition d'un jeu de données jusqu'au traitement des résultats obtenus suite à la division de l'espace des données en hiérarchie de classes. Ci-dessous les différentes étapes par ordre pour une procédure de classification.

II.3.1 Représentation des données

Consiste à prendre connaissance de l'espace des données. Ceci implique la prise en compte du nombre d'objets, du nombre, du type et des échelles de variation des attributs, du nombre de classes (si possible). Cette étape peut également comprendre deux phases de traitement, bien que celles-ci ne soient pas obligatoires :

- Sélection des attributs: vise à déterminer des attributs qui seraient déterminants pour la formation des classes, parmi les attributs existants.

- Extraction des attributs: vise à déterminer de nouveaux attributs pertinents, par application de transformations sur ceux existants.

II.3.2 Définition d'une mesure de distance entre les objets et d'un lien entre les classes

Différentes mesures de distances et de liens sont possibles comme nous avons vu précédemment. Il est alors nécessaire de choisir une mesure particulière pour chacun, convenable au jeu de données.

II.3.3 Classification

Correspond à la phase de construction d'un arbre hiérarchique et mène à l'obtention d'un dendrogramme de la base de données aussi pertinente que possible.

II.3.4 Evaluation de la qualité des résultats

Sur la base de différentes mesures, il est possible d'évaluer la qualité des résultats obtenus.

II.4 Méthodes de classification hiérarchiques

Le fondement de la classification hiérarchique est de créer une hiérarchie de classes. À la racine de l'arbre est associé un unique classe contenant l'ensemble des objets de la base, puis plus on descend dans l'arbre, plus les classes sont spécifiques à un certain groupe d'objets considérés comme similaires. La sortie d'une méthode hiérarchique n'est donc pas directement une partition de l'espace des données, mais un arbre de partitions successives appelé dendrogramme. Un exemple de

dendrogramme est présenté à la figure II.9. L'axe horizontal correspond aux objets tandis que l'axe vertical indique la dis similarité entre les différents niveaux (ou leur similarité selon le choix du sens de l'axe).

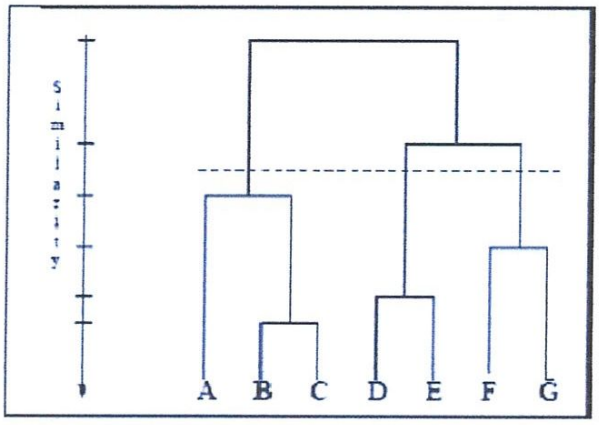


Figure II.9: Exemple de dendrogramme

Afin de former une telle hiérarchie de classes, il existe deux méthodes principales :

1. la méthode ascendante, démarrant avec autant de classes que d'objets initiaux dans la base, puis fusionnant successivement les classes considérées comme les plus similaires, jusqu'à ce que tous les objets soient réunis dans une unique classe stockée à la racine de la hiérarchie formée ;
2. la méthode descendante, démarrant avec un unique classe contenant l'ensemble des objets de la base, puis divisant successivement les classes de manière à ce que les classes résultants soient les plus différents possible, et ce jusqu'à obtenir aux feuilles de la hiérarchie autant de classes que d'objets dans la base.

I.4.1 Les méthodes de classification hiérarchique ascendante

Les méthodes de classification hiérarchique ascendante (CHA) partant des individus isolés assimilés à des classes et procédant, à chaque étape, par agrégation des deux classes les plus proches au sens de la norme choisie (paragraphe ci-dessous). Chaque niveau de hiérarchie représente une partition. Un arbre planaire hiérarchique permet de décrire de façon explicite la structure finale de la classification obtenue : "plus les individus se regroupent en bas de l'arbre, plus ils se ressemblent". Lorsque la norme choisie correspond à une mesure de distance entre objets, on parle dans ce cas de classification hiérarchique ascendante simple. Dans le cas où le critère utilisé pour mesurer la ressemblance entre objets est le moment d'inertie d'ordre deux, on définit la classification hiérarchique ascendante sur le critère du moment d'ordre deux.

II.4.2 Les méthodes de classification hiérarchique descendantes

Les méthodes de classification hiérarchique descendantes(CHD) sont des méthodes de classification divisives. Elles partent de l'ensemble des individus et procèdent par divisions successives de classes jusqu'à l'obtention de classes vérifiant certaine règle d'arrêt. On les appelle aussi méthodes dichotomiques. La complexité d'un algorithme descendant est généralement exponentielle, en effet il se base sur l'énumération complète qui évalue toutes les divisions des *n* individus en deux sous-ensembles non vides, soit 2^{n-1}-1 possibilités. Cette stratégie de l'énumération complète, adoptée par [Cavalli-Sforzal, 65] et bien sur difficilement applicable dés que le nombre *n* d'individus est supérieur à 20.

II.4.3 Algorithmes

II.4.3.1 Hiérarchiques ascendantes

Comme nous l'avons vu, les méthodes hiérarchiques ascendantes consistent à rassembler, à chaque étape, les éléments (objets ou classes) les plus similaires au sein d'une même nouvelle classe. Leur schéma général peut être présenté comme :

1. Créer autant de classes qu'il y a d'objets. Pour *N* objets, on aura donc *N* singletons. Définir une valeur seuil de distance (ou dis similarité) au-dessus de laquelle deux éléments ne devront pas être rassemblés.
2. Comparer toutes les paires d'éléments possibles et marquer la paire ayant la plus petite distance (ou dis similarité).
3. Si cette distance est inférieure à la valeur seuil, rassembler les deux éléments dans un même classe et retourner au point 2. Sinon, fin de la procédure.

Il est évident que le choix de la valeur seuil est déterminant pour la sortie puisqu'elle impose finalement le niveau de coupe dans le dendrogramme. Partant du même principe général, les différentes méthodes hiérarchiques ascendantes se distinguent par la manière d'agglomérer les classes, et plus précisément par la façon de déterminer les deux classes les plus similaires à une étape. On parle de critère d'agrégation. Étant donnée une mesure de distance entre deux classes, ci-dessous, nous allons présenter les différentes méthodes de classification hiérarchique ascendantes [Beck, 06].

a) Méthode basée sur lien simple-Single link (SLINK)

Une méthode de classification agglomérative nécessite la définition d'une distance entre deux classes. Une possibilité est celle du lien minimum. La méthode correspondante sera dite « SLINK ». Cette méthode possède plusieurs particularités, qui peuvent être tantôt perçues comme avantages, tantôt comme inconvénients : tout dépend du jeu de données.

- *effet serpent* : tendance à regrouper les classes de proche en proche et à aboutir à des chaînes dans l'espace des données.
- *souplesse* : permet par exemple de détecter et d'isoler des classes concentriques.
- *tendance à produire de grosses classes rapidement* : il suffit en effet que les deux objets les plus proches entre les classes soient effectivement proches (au sens du seuil) pour rassembler les deux groupes, bien que les classes peuvent s'étaler largement dans les autres directions de l'espace.
- *isole mal les classes mal séparés* : deux classes proches qui devraient être distincts sont facilement rapprochées.
- déséquilibre dans la taille des classes.

Globalement, cette méthode produit de bons résultats pour la détection de classes dans un jeu de données contenant des classes non isotropes (allongées), bien séparées et pouvant présenter des classes concentriques ou en chaînes.

b) Méthode basée sur le Lien complet-Complete link (CLINK)

Un autre critère d'agrégation repose sur le lien minimum (complet) pour définir la distance (ou dis similarité) entre deux classes. La méthode correspondante est dite « CLINK » et a les particularités suivantes :

- tendance à produire des groupes de tailles similaires, isotropes.
- tendance à former des petites classes compactes

Cette méthode performera donc en général assez bien sur un jeu de données où les classes à détecter sont isotropes, bien isolés et de tailles équilibrées.

c) Méthode basée sur le Lien moyen de groupe-Group average link (GALINK)

Une autre mesure de distance (ou dis similarité) entre deux classes sur laquelle on peut se baser est celle du lien moyen de groupe, encore dite lien des centroïdes. La méthode correspondante est dite « GALINK » et possède des particularités intermédiaires aux deux méthodes précédentes.

d) Méthode basée sur le lien moyen-Average link (ALINK)

Le critère d'agrégation peut encore se baser sur le lien moyen. La méthode correspondante est alors dite « ALINK » et ses particularités sont :

- tendance à des classes plus uniformes à partir du moment où tous les objets participent à la mesure de la distance (ou dis similarité) entre les deux classes. chaque objet est en moyenne plus proche de son groupe que de tout autre groupe.
- tendance à détecter des groupes isotropes.
- les tailles des groupes peuvent être différentes.

De manière générale, cette méthode produira de bons résultats lorsque les groupes à détecter sont plus ou moins isotropes et peuvent avoir des tailles différentes.

e) Méthode de WARD

Cette méthode ne se base pas directement sur une mesure de distance (ou dis similarité) entre classes, mais fait reposer le critère d'agrégation sur la notion de variance. Ainsi, à chaque étape, on rassemblera la paire de groupes produisant la plus petite variance dans le groupe obtenu par agglomération de la paire. Cette méthode est coûteuse en temps de calcul (évaluation des combinaisons possibles et de leur variance), mais produit en général de bons résultats, essentiellement lorsqu'il est nécessaire de détecter des classes sphériques.

II.4.3.2 Hiérarchiques descendantes

Les méthodes hiérarchiques descendantes déterminent, à chaque étape, le groupe courant le moins homogène et le splittent en deux sous-groupes. Leur algorithme général est le suivant :

1. Rassembler tous les objets dans une même classe. Définir une valeur seuil de distance (ou de dis similarité) au-dessus de laquelle deux objets ne pourront pas être considérés comme appartenant à un même groupe.
2. Comparer tous les objets deux à deux dans chaque classe et marquer la paire d'objets ayant la plus grande distance (ou dis similarité).
3. Si cette distance (ou dis similarité) est supérieure à la valeur seuil, splitter le classe correspondant en deux et retourner au point 2. Sinon, fin de la procédure.

II.5 Conclusion

Les méthodes ascendantes démarrent avec les objets présentés dans la base de données, qui est ensuite regroupés à chaque étape, alors que les méthodes descendantes démarrent avec l'ensemble complet, qui est ensuite divisé à chaque étape. Une fois la hiérarchie formée, une étape de validation est nécessaire. Une autre étape est ajoutée pour affiner le résultat fourni à l'utilisateur. Il s'agit alors de déterminer le niveau de coupure le plus approprié à appliquer dans l'arbre pour un regroupement des données aussi pertinent que possible. Ceci revient à couper l'arbre au moment où i'on commence à rassembler des éléments forts dissimilaires.

Chapitre 03

Evaluation des méthodes hiérarchique

III .1 Introduction

Après la classification des objets sous forme d'un arbre hiérarchique de classes, on peut maintenant vérifier est-ce que les hauteurs des liens entre les groupes dans l'arbre hiérarchique reflètent mieux ou non les distances originales entres les objets. En outre, on peut étudier la possibilité d'existence des divisions naturelles des liens.

III.2 Indices de validation

Il y a deux mesures qui accomplies ces deux tâches :

- le coefficient de corrélation 'cophnetic'
- et le coefficient d'inconsistance [Hogg, 87].

III.2.1 Le coefficient de corrélation

On note *Z* la matrice du lien entre les objets et *Y* un vecteur de similarité entre les objets, *c* est une valeur qui mesure la corrélation cophenetic entre *Z* et *Y*.

c calcule le coefficient de corrélation 'cophenetic' de l'arbre hiérarchique représenté par Z. Y contient les distances entre touts les paires d'objets employées pour construire Z. Z est une matrice de dimension (n-1) par 3, avec l'information de distance dans la troisième colonne. Y est un vecteur de dimension n*(n-1)/2.

La corrélation 'cophenetic' pour un arbre de classes est définie comme le coefficient de corrélation cophenetic linéaire entre les distances obtenues dans l'arbre (les hauteurs des liens), et les distances originales utilisées pour construire l'arbre. Ainsi, c'est une mesure qui reflète à quel point l'arbre représente les dissimilarités entre les observations.

La distance cophenetic *D* entre deux observations est représentée dans un dendrogramme par la hauteur du lien auquel ces deux observations sont liées pour la première fois. Cette hauteur est la distance entre les deux sous-classes qui sont fusionnées par ce lien.

La valeur *c*, est le coefficient de corrélation cophenetic ; elle devrait être très proche de 1 pour une solution de bonne qualité.

La corrélation cophenetic c entre D et Y est définie par l'équation suivante :

$$c = \frac{\sum_i (Y_i - y)(D_i - d)}{\sqrt{\sum_i (Y_i - y)^2 \sum_i (D_i - d)^2}}$$

Où :

- *Y_i : est la distance entre deux objets quelconque.*
- *Di : est la distance cophenetic entre deux objets quelconque*
- *y et d : sont respectivement les moyennes de Y et D.*

III.2.2 Coefficient d'inconsistance

Un moyen pour déterminer les divisions naturelles des classes dans un ensemble de données est de comparer la hauteur de chaque lien dans l'arbre de classes avec les hauteurs voisines des liens ci-dessous dans l'arbre.

L'existence d'un lien qui a approximativement la même hauteur que les liens ci-dessous, indique qu'il n'existe pas des divisions distinctes entre les objets groupés à ce niveau hiérarchique. Ces liens montrent un niveau élevé de consistance. Parce que la distance entre les objets qui sont reliés est approximativement identique aux distances entre les objets qu'ils contiennent.

D'autre part, l'existence d'un lien dont la hauteur diffère considérablement de la hauteur des liens ci-dessous indique que les objets groupés à ce niveau sont plus éloignés. Ce lien est considéré comme inconsistant avec les liens en dessous.

Dans l'analyse des classes, les liens inconsistants peuvent indiquer la bordure d'une division naturelle dans un ensemble des données. A cet effet, le coefficient d'inconsistance représente le niveau de coupure dans l'arbre hiérarchique pour extraire finalement les classes naturelles existantes dans la l'ensemble de données [Hogg, 87].

Le dendrogramme suivant, illustre des liens inconsistants. Il montre comment les objets dans le dendrogramme se descendent en trois groupes qui sont connectés par des liens de niveau très élevé dans l'arbre. Ces liens sont inconsistants en comparaison avec ceux au-dessous dans la hiérarchie.

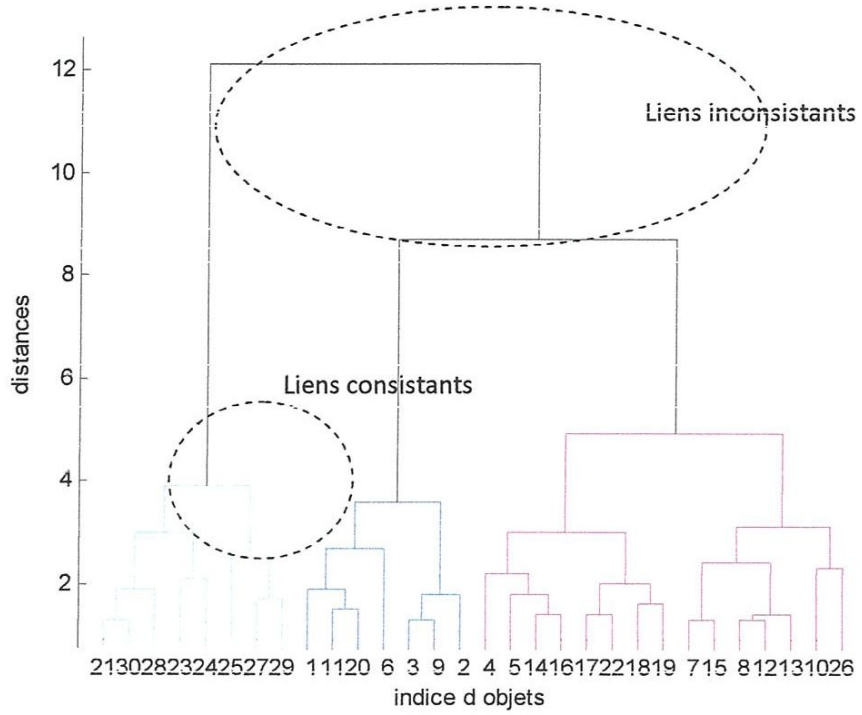


Figure III.1 : Exemple d'un dendrogramme présente des liens inconsistants et consistants

III.2.3 Consistance relative

La consistance relative de chaque lien dans un arbre hiérarchique peut être calculée et exprimée comme le coefficient d'inconsistance de chaque lien. Cette valeur compare la hauteur de ce lien avec la moyenne des hauteurs des liens au-dessous. Les liens qui joignent des classes distinctes ont un coefficient d'inconsistance élevé tendis que les liens qui joignent des classes similaires ont un coefficient d'inconsistance faible.

III.2.4 Calcul des coefficients d'inconsistance

Le coefficient d'inconsistance caractérise chaque lien dans une hiérarchie en comparant sa hauteur avec la moyenne des hauteurs des autres liens du même niveau dans l'hiérarchie. Une valeur élevée de ce coefficient, indique que les objets qui sont connectés par ce lien sont moins semblables [Jain, 88].

Pour chaque lien k, le coefficient d'inconsistance est calculé comme la suite :

$$Y(k,4) = (Y(k,3) - Y(k,1))/Y(k,2)$$

Où :

Y, est une matrice de dimension (*n*-1) par 4 composée comme la suite :

- *Y(k,1) :* Moyenne des hauteurs de tous les liens inclus dans le calcul.
- *Y(k,2) :* Écart type de tous les liens inclus dans le calcul.
- *Y(k,3) :* Nombre de liens inclus dans le calcul.
- *Y(k,4) :* Coefficient d'inconsistance.

Les derniers nœuds dans le dendrogramme (feuilles), ont des coefficients d'inconsistance nuls.

III.5 conclusion

Le *coefficient de corrélation* cophenetic compare les distances 'cophenetic' avec les distances réelles. Si la classification est valide, alors les liens entre les objets dans l'arbre devraient avoir une corrélation cophenetic forte avec les distances entre les objets dans le vecteur de distance. Le coefficient de corrélation cophenetic se résulte de calcule de la corrélation cophenetic entre ces deux ensembles. Plus la valeur du coefficient de corrélation 'cophenetic' est proche de 1, plus le dendrogramme reflète mieux la structuration des données.

On peut utiliser le coefficient de corrélation cophenetic pour comparer les résultats de la classification du même jeu de données en utilisant différentes méthodes de calcul de distance entres les objets avec différents liens entres les classes.

Un moyen pour déterminer les divisions naturelles des classes dans un ensemble de données est de comparer la hauteur de chaque lien dans l'arbre de classes avec les hauteurs voisines des liens ci-dessous dans l'arbre. Le coefficient d'inconsistance le plus élevé représente le niveau de coupure de l'arbre hiérarchique pour extraire finalement les classes naturelles existantes dans la l'ensemble de données.

Chapitre 04

Expérimentation

Pour valider les résultats de la classification hiérarchique, nous avons implanté sur MATLAB les algorithmes de classification hiérarchique ascendante suivants :

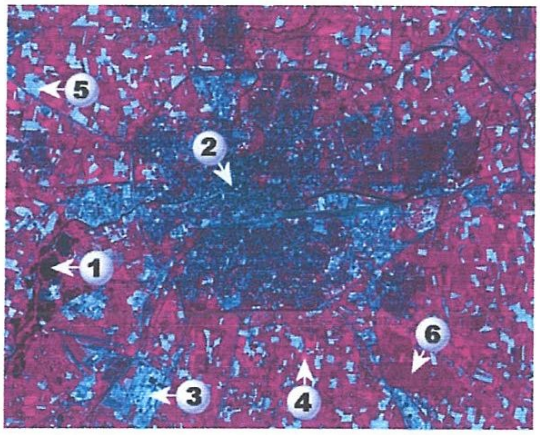
- Algorithme basé sur le lien simple (Single link: SLINK) ;
- Algorithme basé sur le lien moyen (Average link : ALINK) ;
- Algorithme basé sur lien complet (Complet link: CLINK) ;
- Algorithme basé sur le lien moyenne de groupe (Group Average link: GALINK) ;
- Algorithme de Ward.

Nous allons tester chaque algorithme avec différentes mesures de distance entre objets (euclidienne, Mahalanobis, Cityblock et Minkowski) sur deux base de données : une image satellitaire de rennes et la base de données iRiS.

IV.1 Description des bases de données utilisées

[1] La base de données Iris est constituée de 150 fleurs décrites par 4 variables (longueur et largeur de sépales, et de pétales), le nombre de classes est égal à 3, les objets sont uniformément répartis en trois classes, les classes 2 et 3 sont facilement séparables de la classe 1, mais difficilement séparables entre elles.

[2] L'image ci-dessous représente la ville de rennes, elle contient 6 régions, chaque pixels dans l'image est caractérisée par 5 attributs (coordonnées x, y et composante RVB).l



1 → Etangs d'Apigné , 2 → Centre ville , 3 → Usine Citroen , 4 → Champs cultivés ou prairies , 5 → Champs récoltés ou labourés , 6 → forêts

Figure IV.1 : Image rennes captée par satellite spot

IV.3 Algorithme

Quelque soit la méthode du lien utilisée, un algorithme de classification hiérarchique ascendante consiste à :

1. *Trouver la similarité entre chaque paire d'objets dans jeu de données* : dans cette étape, en calcule la distance entre les objets pour déterminer la proximité des objets les uns par apport aux autres. .

2. *Grouper les objets dans un arbre de classe hiérarchique binaire* : dans cette étape nous créerons un lien entre paires d'objets qui sont proches en utilisons un des liens cités précédemment.

3. *Déterminer le niveau de coupure de l'arbre hiérarchique* : dans cette étape en coupe les branches de l'arbre, et on assigne tout les objets au-dessous de chaque branche a une seule classe ce qui permet de créer une partition de données.

iV .4 Présentation des résultats sur MATLAB

Un arbre hiérarchique dans le MATLAB se présente sous forme d'un dendrogramme. il est constitué autour de plusieurs linges en forme de U reliant des objets dans une arborescence hiérarchique. La hauteur de chaque U représente la distance entre les deux objets connectés. Si il y avait 30 point ou moins de données dans l'ensemble de données d'origine, chaque feuille dans le Dendrogramme correspond à un point de données. S'il y avait plus de 30 point de données, l'arbre complet peut paraitre encombré, et le dendrogramme se mélange au niveau des branches inférieures, de sorte que quelques feuilles dans la parcelle correspondent à plus d'un point de données.

IV.4.1 Exemple typique

Dans cet exemple, on considère une base de données simple qui comporte 5 personnes. Chaque personne présente respectivement deux caractéristiques : le poids et l'âge.

Personne 1 (64,35) ⟶ Object 1

Personne2 (58,30) ⟶ Object 2

Personne3 (61,26) ⟶ Object 3

Personne 4 (63,21) ⟶ Object 4

Personne 5 (63, 17) ⟶ Object 5

On essaye d'organiser ces personnes dans un arbre hiérarchique pour illustrer le degré de liaison entre eux. La base de données peut se mettre sous forme de cette matrice :

$$X = \begin{bmatrix} 63 & 35 \\ 58 & 17 \\ 61 & 21 \\ 63 & 30 \\ 63 & 26 \end{bmatrix}$$

- ***Etape 1*** *: calcul de la distance entre les objets*

On calcule la distance entre l'objet 1 et objet 2, l'objet 1 et objet 3, et ainsi de suite jusqu'à ce que les distances entre toutes les paires ont été calculées. La figure suivante présente ces objets graphiquement. La distance euclidienne entre l'objet 2 et objet 3 est indiquée pour illustrer une interprétation de la distance.

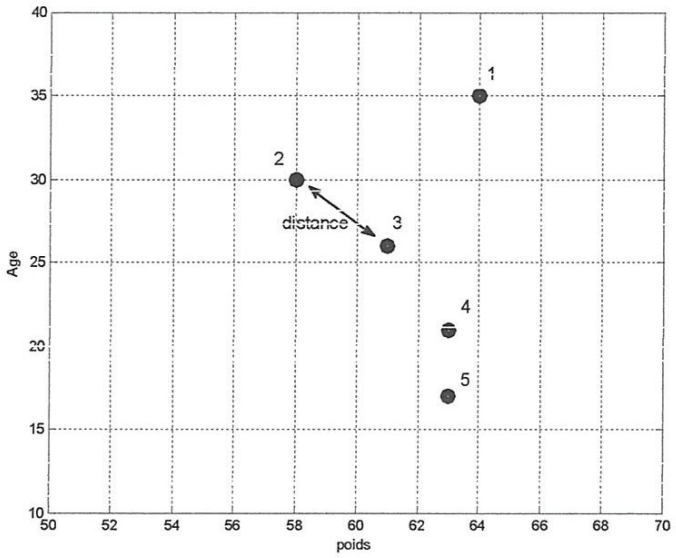


Figure IV.2: Mesure de distance euclidienne entre les objets 2 et 3.

Le vecteur Y contient toutes les mesures de distance entre les différentes combinaisons possibles :

Y= [7.81 9.49 14.04 18.03 5.00 10.30 13.93 5.39 9.22 4.00].

- ***Etape 2 :*** *Calcul du lien entre les groupes (construction du dendrogramme)*

Une fois la proximité entre les objets dans le jeu de données a été calculée, on peut

déterminer comment les objets dans le jeu de données devraient être regroupés en classes, en utilisant l'un des liens. Ce dernier génère un arbre de classification hiérarchique, présenté par une matrice Z.

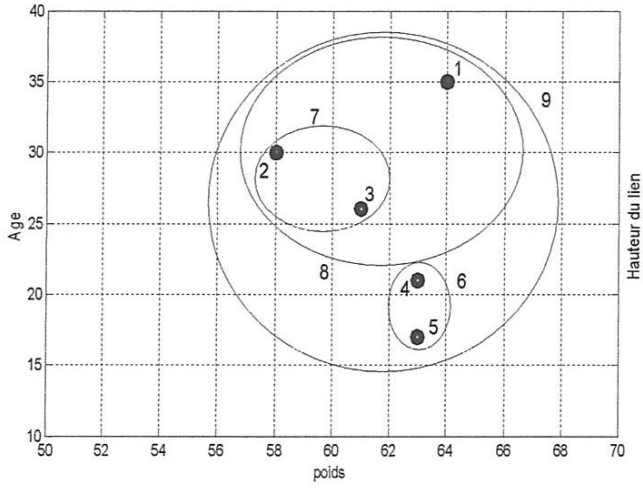
Z=

4.00	5.00	4.00
2.00	3.00	5.00
1.00	7.00	8.65
6.00	8.00	11.82

Dans cette matrice, chaque ligne identifie un lien entre les objets ou classes. Les deux premières colonnes identifient les objets qui ont été liées. La troisième colonne contient la distance entre ces objets. Pour notre exemple, le lien commence par le regroupement d'objets 4 et 5, qui sont les plus proche (valeur de distance = 4), ensuite il continu par le regroupement des objets 2 et 3, qui ont une valeur de distance de 5.

Les données initiales ne comportent que cinq objets, donc l'objet 6 est la nouvelle classe binaire créé par le regroupement d'objets 4 et 5. Lorsque le lien regroupe deux objets (classes) dans une nouvelle classe, il doit assigner à la nouvelle classe un indice en commençant par la valeur n +1 où n est le nombre d'objets dans les données initiales. (Valeur: 1 à n sont déjà utilisées par l'ensemble des données originale). Du même, l'objet 7 est le groupe formé par le regroupement d'objets 2 et 3. Le vecteur de distance Y contient les distances entre les objets d'origine 1 à 5. Mais le lien doit également être capable de déterminer les distances entre les classes crées ou entre un objet et une classe (entre 1 et 7 par exemple).

Pour la classe finale, le lien regroupe l'objet 6 (la nouvelle classe composée d'objets 4 et 5), avec l'objet 8. Les deux figures suivantes illustrent graphiquement la manière dont le lien regroupe les objets dans une hiérarchie de classes.



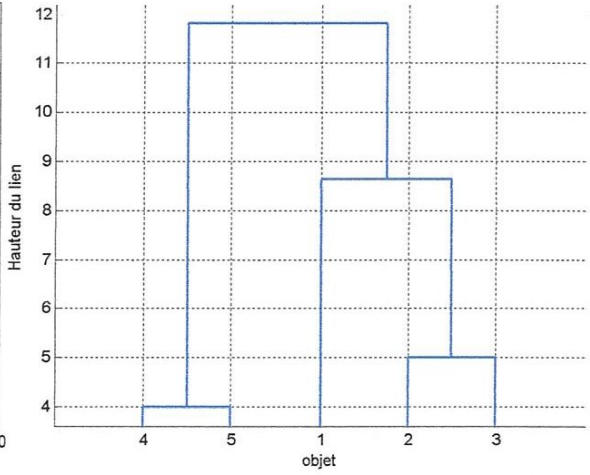


Figure IV.3 : Deux manières de représentation d'un arbre hiérarchique avec la méthode ALINK.

– ***Etape 3*** *: validation de l'arbre hiérarchique*

Après la liaison de touts les objets dans un dendrogramme, on passe à l'étape de validation. Deux mesures sont disponibles :

- ***Coefficient de corréiation cophenetic***
- ***Coefficient d'inconsistance***

a. Vérification de corrélation cophenetic

Une valeur du coefficient de corrélation cophenetic proche de 1 indique une bonne organisation des données dans le dendrogramme. A cet effet, on doit exécuter les 5 algorithmes avec différentes mesures des distances. Si on applique l'algorithme ALINK) avec une distance de Mahalanobis, on trouve une valeur de 0.85 qui présente le meilleur résultat. Le dendrogramme équivalent est présenté dans la figure **IV.4.**

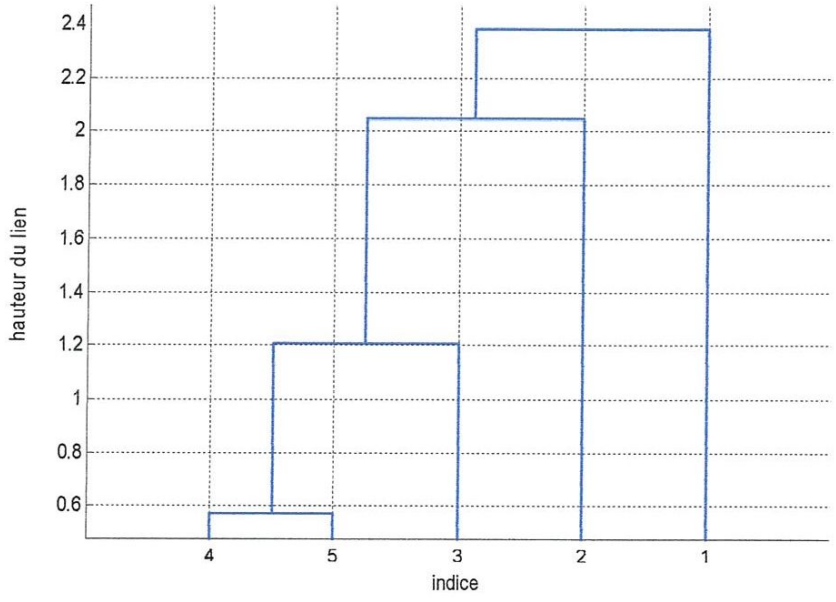


Figure IV.4 : Dendrogramme obtenu avec la méthode ALINK et distance de Mahalanobis.

La nouvelle matrice Z est donnée comme la suite :

Z= 4.00 5.00 0.57

3.00 6.00 1.21

2.00 7.00 2.05

1.00 8.00 2.38

b . Vérification de l'inconsistance

Après avoir choisir la méthode du lien ainsi que la distance les plus convenable, on cherche maintenant à calculer l'inconsistance de chaque lien dans le dendrogramme. Le lien qui présente une inconsistance élevée joigne deux groupes distincts. Donc, cette mesure nous permet de trouver le niveau de coupure de l'arbre pour crier finalement une partition de l'ensemble de données en des classes naturelles.

Pour chaque lien, k, le coefficient d'inconsistance est calculé comme la suite :

$$I(k,4) = \frac{Z(k,3) - I(k,1)}{I(k,2)}$$

En appliquant la méthode ALINK avec une mesure de distance de Mahalanobis, on trouve la matrice d'inconsistance suivante :

I=

0.57 0 1.00 0

0.89 0.45 2.00 0.71

1.28 0.74 3.00 1.04

1.55 0.82 4.00 1.01

La première colonne dans la matrice d'inconsistance définit la moyenne des hauteurs de tous les liens ci-dessous inclus dans le calcul. La deuxième donne l'écart type de tous les liens inclus dans le calcul qui sont définies dans la troisième. Et finalement, l'inconsistance de chaque lien est définie dans la quatrième colonne. Chaque ligne dans la matrice d'inconsistance présente un lien (4 lien).

Par exemple la première ligne exprime le premier lien dans le dendrogramme (entre l'objet 4 et 5) qui a une hauteur de 0.57 (d'après Z), l'écart type égale à 0, une lien qui entre dans le calcul et donc une inconsistance nulle.

Prenons maintenant la dernière ligne : selon Z la hauteur du lien est de 2.38 qui est entre 1 et 8, selon le dendrogramme, il apparaît 3 lien en dessous de cet lien, donc 4 liens qui entre dans le calcul, la moyenne sera :

I(4,1)= [Z(1,3)+Z(2,3)+Z(3,3)+Z(4,3)]/4=(0.57+1.21+ 2.05+2.38)/4= 1.55 ;

Donc l'inconsistance égale à :

$$I(4,4) = \frac{Z(4,3) - I(4,1)}{I(4,2)}$$

$$= (0.57\text{-}1.55)/0.97$$
$$= -1.01$$

c. Choix de niveau de coupure

Après la reconnaissance de l'inconsistance de chaque lien dans l'arbre, on doit choisir maintenant, le lien qui joigne les deux groupes les plus inconsistants. Une valeur de coupure entre 1.21 et 2.50 répond à cette exigence. En fin, on assigne chaque objet au groupe équivalent. Le dendrogramme ci-dessous montre la partition finale des 5 objets de notre exemple en trois classes.

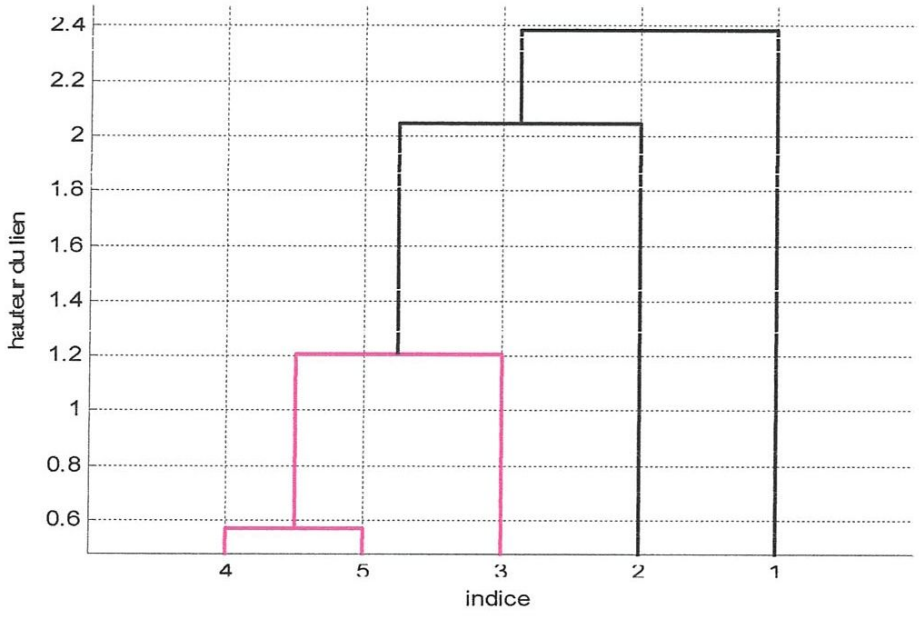


Figure IV.5 : Partition finale en trois classes pour une coupure c=1.6.

A l'aide du dendrogramme, on peut maintenant assigner chaque objet à la classe équivalente comme la suite :

Objet 1 ⟶ Classe 1

Objet 2⟶ Classe 2

Objet 3⟶ Classe 3

Objet 4 ⟶ Classe 3

Objet 5 ⟶ Classe 3

IV. 5 Résultats de la classification de la base de données iris

On recommence la même procédure précédente pour valider les résultats avec la base de données Iris. Le tableau IV. 1 montre les résultats.

Lien	Distance	Coefficient de corrélation cophenetic	Erreur Totale(%)	Coefficient d'inconsistance du lien de coupure
SLINK	Euclidienne	0.86	32.00	1.46
	Mahalanobis	0.64	66.66	0.84
	Cityblock	0.85	32.66	1.13
	Minkowski	0.86	32.00	1.46
ALINK	Euclidienne	**0.87**	**09.33**	3.10
	Mahalanobis	**0.67**	**12.66**	1.55
	Cityblock	**0.86**	**10.00**	3.20
	Minkowski	**0.87**	**09.33**	3.10
CLINK	Euclidienne	0.72	99.33	4.21
	Mahalanobis	0.47	75.33	2.92
	Cityblock	0.68	66.66	5.20
	Minkowski	0.72	99.33	4.21
GALINK	Euclidienne	**0.87**	**09.33**	2.76
	Mahalanobis	0.65	65.33	1.21
	Cityblock	0.86	100.00	3.42
	Minkowski	**0.87**	**09.33**	2.76
Ward	Euclidienne	0.87	10.66	5.75
	Mahalanobis	0.52	18.00	4.08
	Cityblock	0.86	11.33	5.62
	Minkowski	0.87	10.66	5.75

Tableau IV.1 : Taux d'erreur, coefficient de corrélation cophenetic et coefficient d'inconsistance équivalents aux différentes méthodes du lien pour chaque mesure de distance appliquée sur Iris.

D'après le tableau 4, on peut noter les remarques suivantes :

- Les grandes valeurs du coefficient de corrélation cophenetic correspond aux faibles pourcentages d'erreur.
- La méthode ALINK présente des meilleurs résultats en termes des taux d'erreurs les plus faibles et coefficient de corrélation cophenetic les plus élevés ;
- La distance euclidienne donne le même résultat que celle de Minkowski ;
- La distance Euclidienne montre toujours le meilleur résultat sauf avec CLINK ;
- La distance Euclidienne (ou Minkowski) appliquée sur la méthode ALINK (ou GALINK) présente le meilleur résultat avec le minimum des taux d'erreur égale à 9.33% et le maximum des corrélations cophenetic de 0.87.
- La distance Mahalanobis donne le mauvais résultat avec tous les liens, sauf avec CLINK.
- La méthode de Ward montre une inconsistance élevée.

Les dendrogrammes obtenus avec SLINK- Euclidienne, SLINK- Mahalanobis, ALINK-Euclidienne et GALINK-Euclidienne sont présentés dans la figure IV .6.

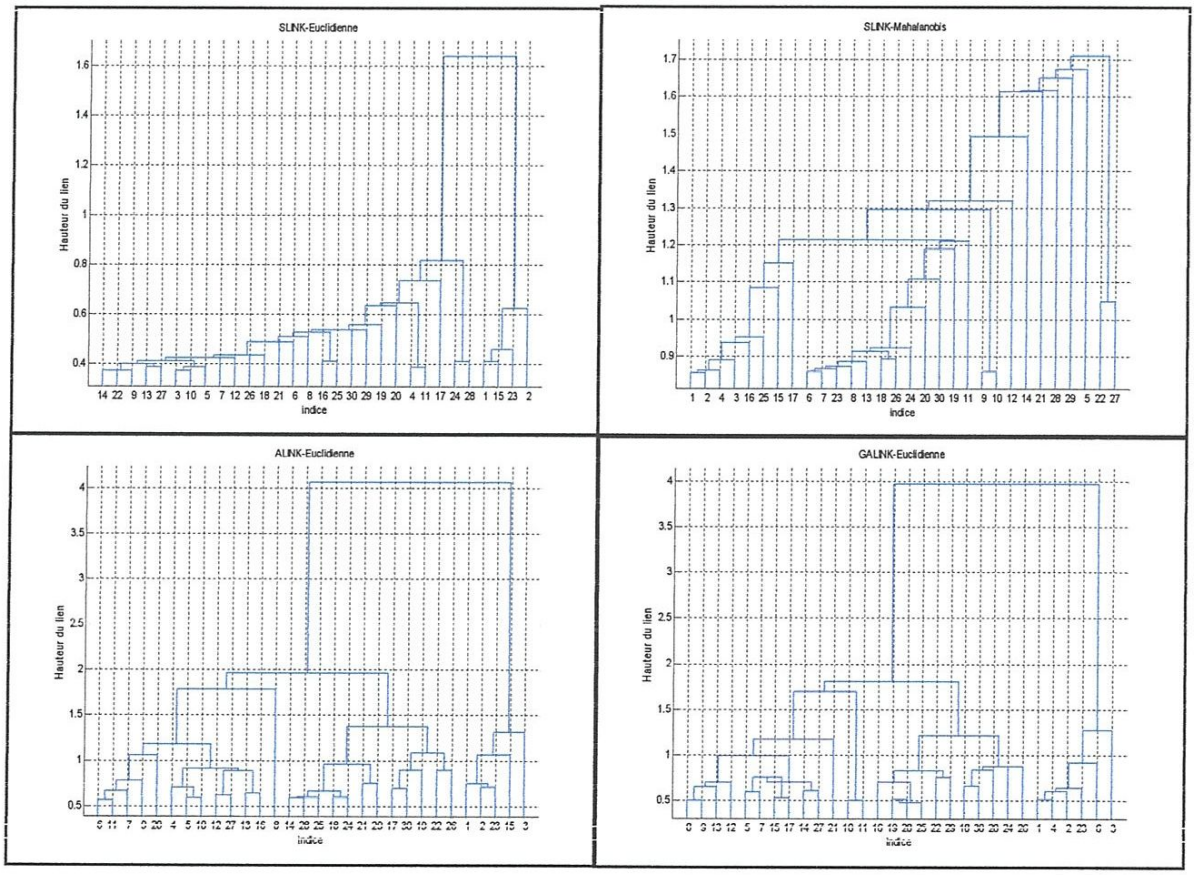


Figure IV .6 : Dendrogrammes obtenus avec SLINK- Euclidienne, SLINK-Mahalanobis, ALINK- Euclidienne et GALINK-Euclidienne appliqués sur Iris.

L'évolution du lien et de son inconsistance correspondante à chaque dendrogramme est présentée dans les figures ci-dessous.

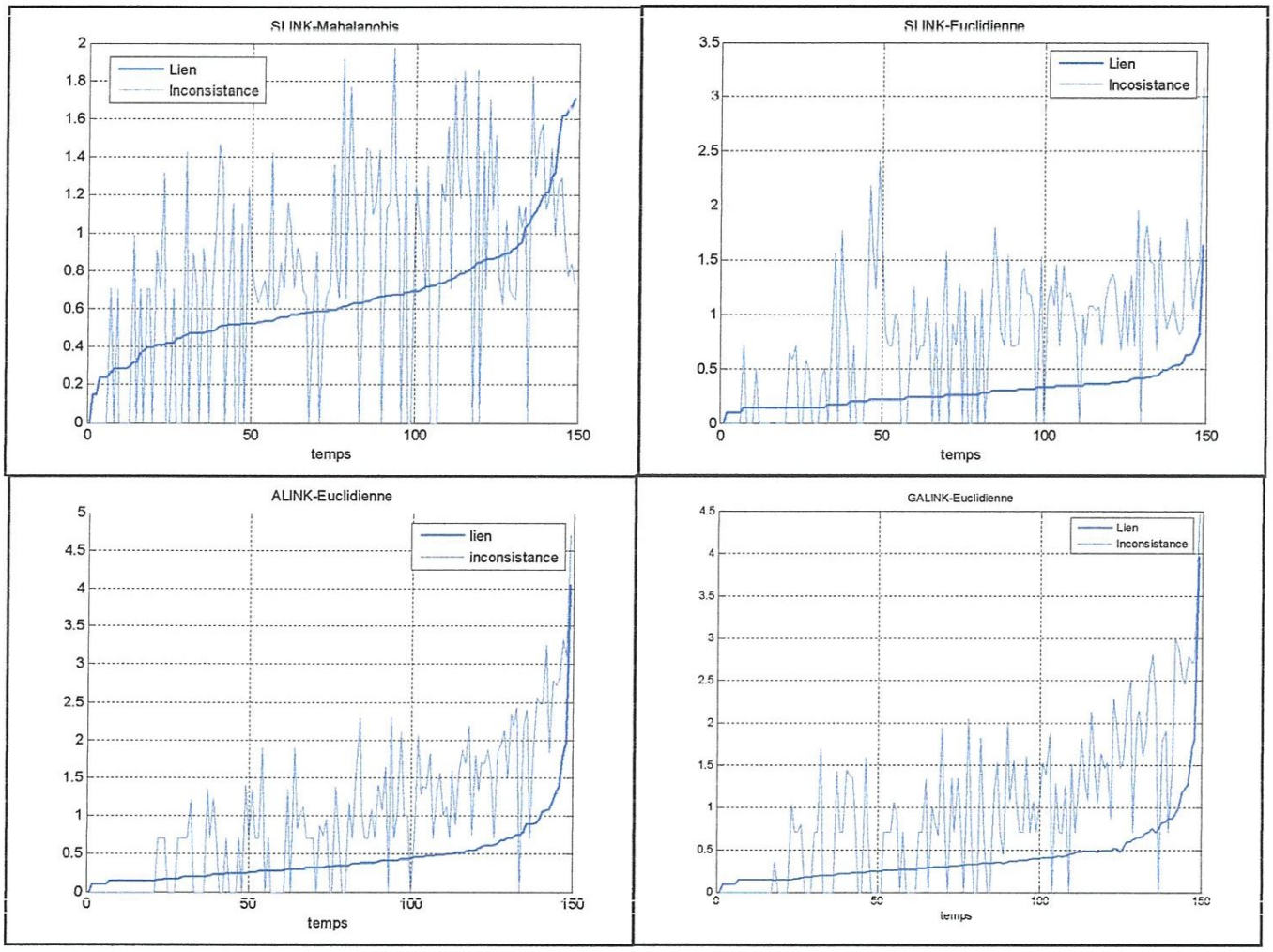


Figure iV. 7 : Evolution du lien et de l'inconsistance de iris.

En analysant les courbes, on remarque :

- Les quatre courbes montrent la croissance de la hauteur du lien en fonction du temps ;

- Généralement, avec ALINK-Euclidienne et GALINK-Euclidienne l'inconsistance est croissante. Tandis qu'avec SLINK- Euclidienne et SLINK-Mahalanobis, elle n'a pas une allure claire.

Pour un nombre de classe égale à 3, on choisie la hauteur du lien de coupure. Avec ALINK-Euclidienne le dendrogramme donne la partition finale en trois classes comme la suite :

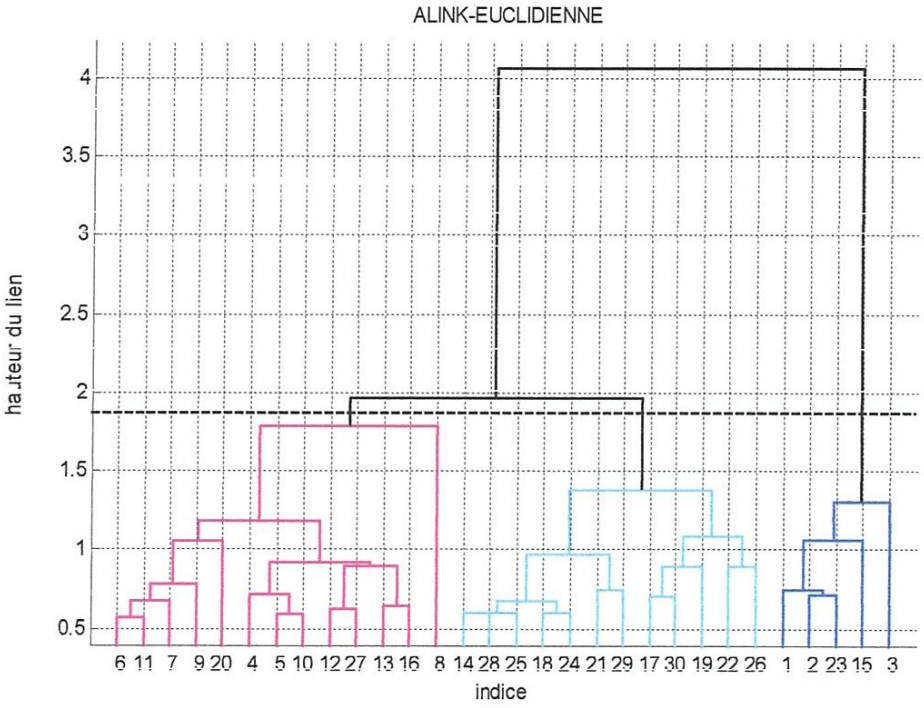


Figure IV.8 : Partition finale en trois classes pour une coupure c=1.9.

Les matrices de confusion obtenues avec SLINK- Mahalanobis, ALINK-Euclidienne et GALINK- Euclidienne sont les suivantes :

1 0 0	0 0 36	0 0 36
49 50 48	0 50 14	0 50 14
0 0 2	50 0 0	50 0 0
SLINK- Mahalanobis Erreur totale :100	ALINK-Euclidienne Erreur totale : 14	GALINK- Euclidienne Erreur totale :14

Conclusion :

D'après les résultats obtenus et les remarque cités auparavant, on peut conclure la pertinence de ALINK (appliqué avec la distance Euclidienne ou de Minkowski) et GALINK (appliqué avec la distance Euclidienne) .

IV. 6 Résultats de la classification de la l'image satellitaire

Ici nous utilisons une partie de l'image satellitaire de la ville de rennes (voir figure IV.1) pour reconnaitre les régions simulé afin de dessiner la carte finale de cette ville. Les résultats sont rapportés dans le tableau suivant :

Lien	Distance	Coefficient de corréiation cophenetic
SLINK	Euclidienne	0.72
	Mahalanobis	0.45
	Cityblock	0.67
	Minkowski	0.72
ALINK	Euclidienne	**0.84**
	Mahalanobis	0.45
	Cityblock	0.67
	Minkowski	0.72
CLINK	Euclidienne	0.78
	Mahalanobis	0.45
	Cityblock	0.67
	Minkowski	0.72
GAL INK	Euclidienne	**0.84**
	Mahalanobis	0.45
	Cityblock	0.67
	Minkowski	0.72
Ward	Euclidienne	**0.84**
	Mahalanobis	0.54
	Cityblock	0.80
	Minkowski	**0.84**

Tableau IV.2 : coefficient de corrélation cophenetic équivalents aux différentes méthodes du lien pour chaque mesure de distance appliquée sur l'image satellitaire.

A la lecture du tableau IV.2, on remarque :

- La méthode Ward présente des coefficients de corrélation cophenetic les plus élevés ; par contre SLINK montre les plus faibles.
- Pour chaque lien, la distance Euclidienne présente toujours le maximum des coefficients de corrélation cophenetic ; tandis que Mahalanobis montre toujours le minimum ;
- Le meilleur résultat est vérifié avec la distance Euclidienne appliquée sur ALINK, GALINK et WARD ;

Les quatre dendrogrammes obtenus avec SLINK-Mahalanobis, ALINK-Euclidienne, GALINK- Euclidienne et Ward- Euclidienne sont les suivants :

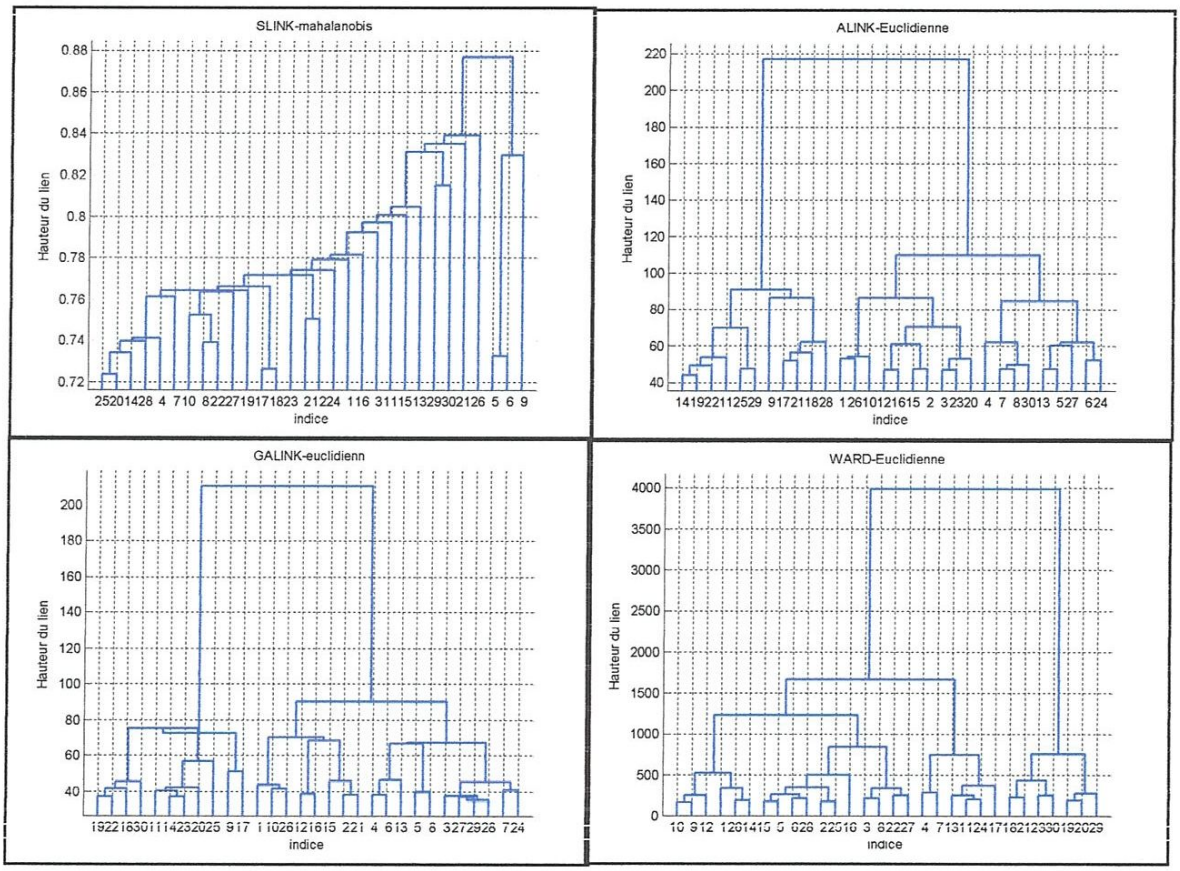


Figure IV .9 : Dendrogrammes obtenus avec SLINK-Mahalanobis, ALINK-Euclidienne, GALINK- Euclidienne et Ward- Euclidienne appliqués sur l'image.

L'évolution du lien et d'inconsistance équivalente à chaque dendrogramme de la figure IV.9 est montrée ci-dessous :

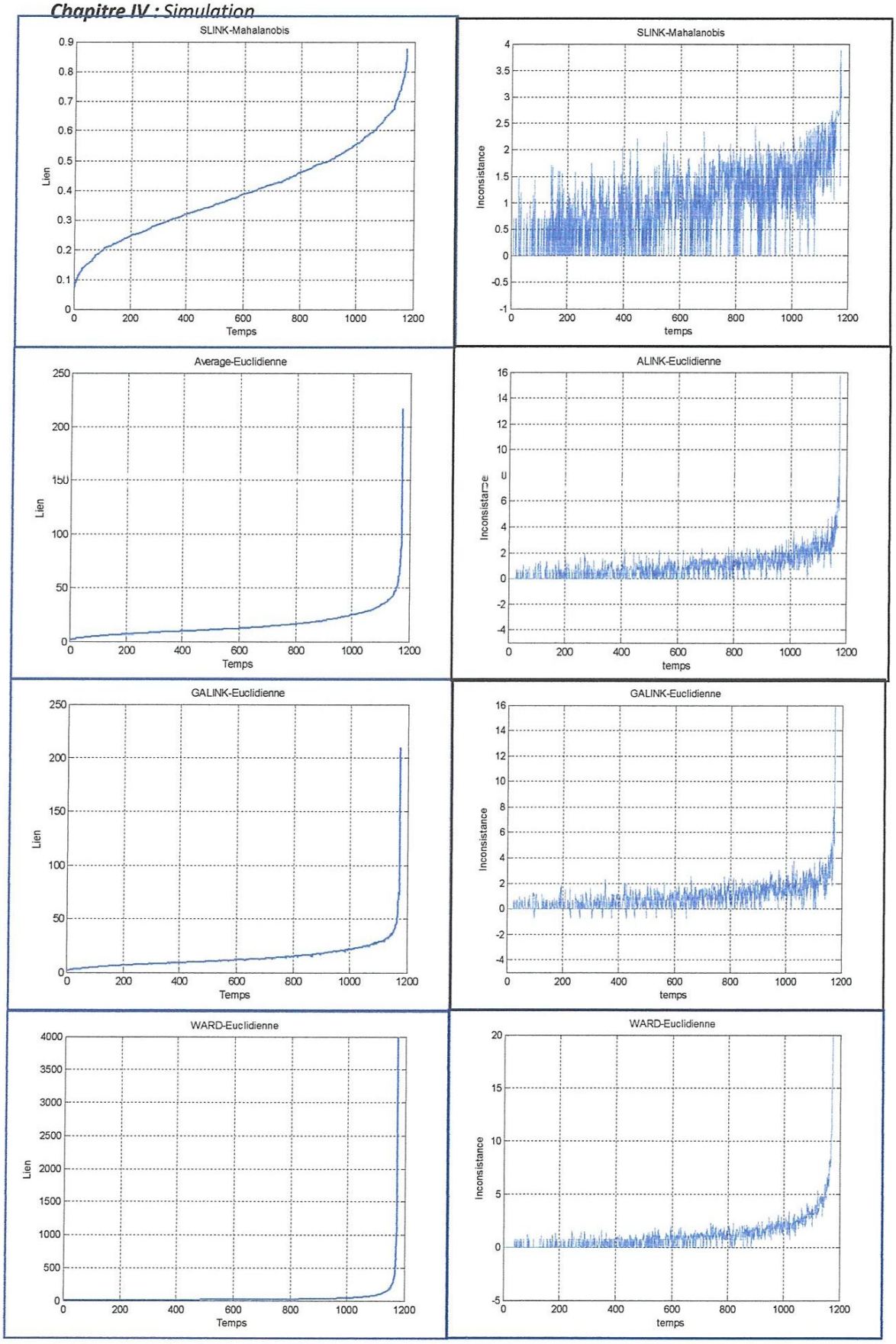


Figure IV. 10 : Evolution du lien et de l'inconsistance de l'image.

En analysant les courbes, on note clairement la manière de variation de l'inconsistance avec SLINK-Mahalanobis. Au contraire, avec ALINK-Euclidienne, GALINK- Euclidienne et Ward- Euclidienne, la variation est croissante est présente une allure semblable. Pour un nombre de classe égale à quatre le dendrogramme, le dendrogramme WARD-Euclidienne donne la partition finale de l'image.

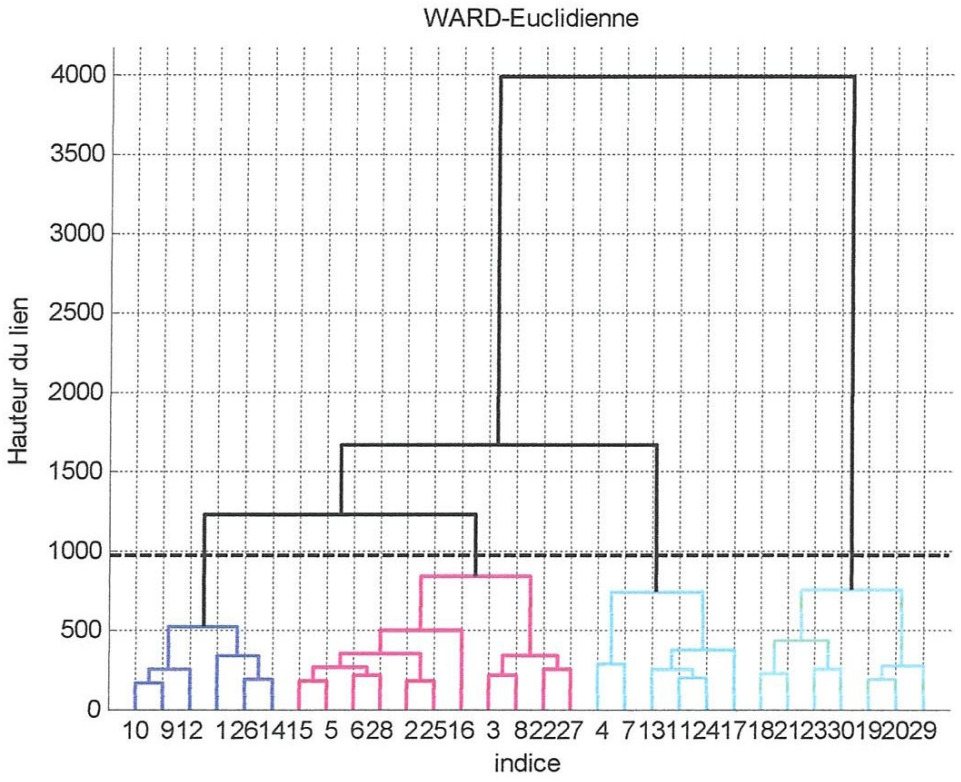


Figure IV.11 : Partition finale en quatre classes pour une coupure c=1000.

Chaque couleur dans le dendrogramme représente une classe qui contient plusieurs indices d'objets. Du même, chaque indice contient un sous- ensemble d'objets (pixels). On peut donc reconstituer l'image à partir de ces pixels pour arriver finalement à définir la carte de cette image.

IV. 7 Interface réalisé par programme MAtlAB :

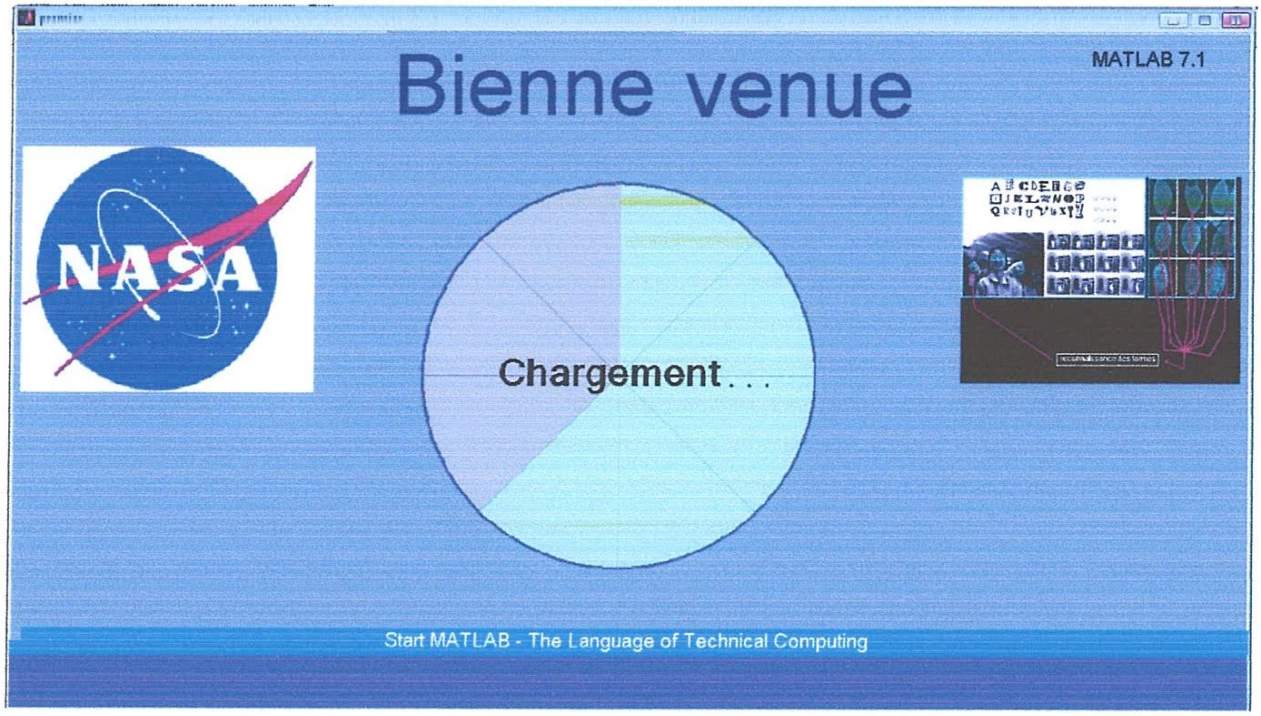


Figure IV.12 : page d'accueille .

Page d'exécution

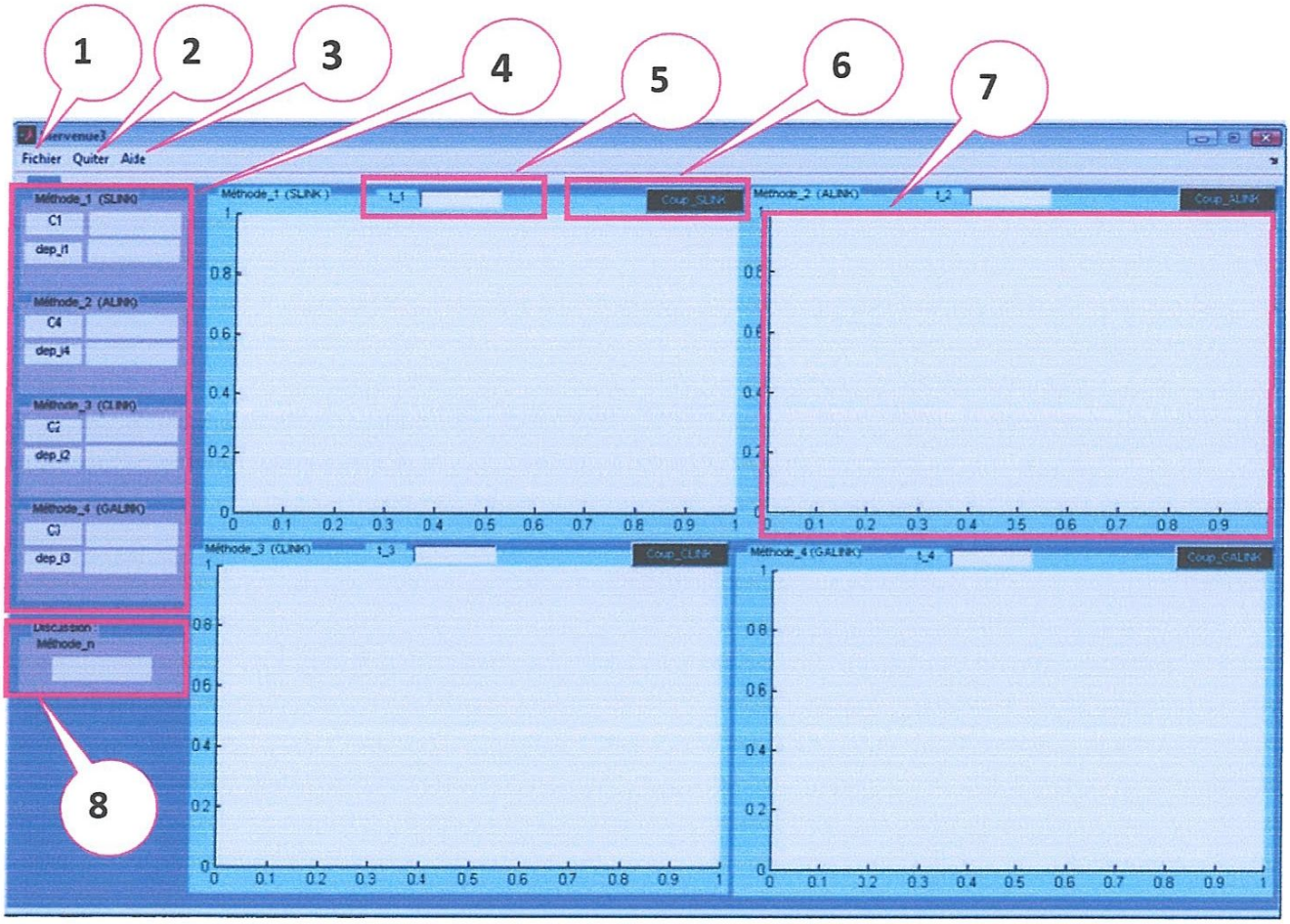


Figure IV.12 : page d'accueille .

1. Fichier (chargement d'image) pour chargées l'image.

2. Pour atteindre le programme.

3. Pour obtenue a information.

4. pour afficher le coefficient de corrélation 'cophnetic 'et le maximum de valeur d'inconsistance.

5. chargement manuelle la valeur de coupure.

6. le bouton de coupure.

7. espace pour affichage (arbre hiérarchique).

8. la décision de meilleure méthode.

IV. 6 Conclusion

Après cette étude, on peut conclure que la méthode ALINK et GALINK donne les meilleurs résultats en appliquant une mesure de distance euclidienne. Les deux mesures de validité montrent leur concordance et tendance à évaluer les résultats avec précision.

Conclusion générale

Dans ce mémoire de fin de d'études, nous avons étudié la validation des algorithmes hiérarchiques ascendants pour mesurer les performances d'un algorithme par rapport à un autre. Comme il n'existe pas un algorithme qui est absolument performant, toujours une comparaison entre autres algorithme est nécessaire pour trouver le meilleur. A cet effet, on fait recours aux méthodes de validation. Parmi ces dernières, nous avons présenté deux mesures differentes ; le coefficient de corrélation cophnetic et le coefficient d'inconsistance

Après la création de l'arbre hiérarchique de classes binaires, le coefficient de corrélation cophnetic vérifie est ce que les objets sont bien structurée dans le dendrogramme par la mesure de la corrélation entre la hauteur du lien et les distance entre tous les paires d'objets sous ce lien. La valeur de ce coefficient proche de 1 indique une bonne organisation des données dans le dendrogramme, c .à..d une bonne classification de ces dernières.

Les testes expérimentaux réalisés sur les deux bases de données montrent d'une part, la robustesse des ALINK et GALINK appliqués sur une mesure de distance Euclidienne pour la création des classes naturelles. Et d'autre part, la fiabilité de la corrélation cophnetic d'exprimer le résultat de la classification. Du même, l'inconsistance démontre l'existence ou non des liens inconsistants dans l'arbre hiérarchique. La simulation démontre aussi la concordance des deux mesures de validité d'exprimer les résultats.

En fin, l'inconsistance peut identifier les liens où les similarités entre les objets changent brusquement, ceci nous permet de couper l'arbre pour partitionner finalement les données en classes naturelles distincts bien-séparés.

En perspective, nous envisagerons de chercher d'autres méthodes hiérarchiques qui améliorent également le résultat de la classification.

Bibliographie

[Duda, 73] Duda, R., Hart, P., 1973. 'Pattern Classification and Scene Analysis.Wiley' NewYork, USA.

[Andrw, 99] Andrew Webb,' statistical pattern recognition', Arnold edition, 1999.

[Jain, 88] Jain, A., and R. Dubes. *Algorithms for Clustering Data*. Upper Saddle River, NJ: Prentice-Hall, 1988.

[Zahn, 71] Zahn, C. T. "Graph-theoretical methods for detecting and describing Gestalt clusters." *IEEE Transactions on Computers*. Vol. C-20, Issue 1, 1971, pp. 68–86.

[Beck, 06] Nicolas Beck, Application de méthodes de clustering traditionnelles et extension au cadre multicritère, mémoire d'Ingénieur, Université libre de Bruxelle, 2006.

[Cavalli, 65] Edwards et Cavalli-SforzaL.L(1965), A method for cluster analysis, Biometrics 21, pp.362-375.

[Hogg, 87] R. V., and J. Ledolter, Engineering Statistics, MacMillan, 1987.

« Classification Automatique et reconnaissance de formes »fichier internet :

http:// www.inria.fr/rapportsactivite/RA95/colorec.html.

http://www.mathworks.com/products/fuzzylogic.html

http://www.mathworks.com/products/fuzzylogic.html