



UNIVERSITE MOHAMED BOUDIAF - M'SILA
FACULTE DES MATHÉMATIQUES ET
DE L'INFORMATIQUE



DEPARTEMENT D'INFORMATIQUE

MEMOIRE DE FIN D'ETUDE

Présenté pour l'obtention du diplôme de **MASTER**

Domaine : Mathématiques et Informatique

Filière : Informatique

Option : Systèmes d'Information et Génie Logiciel

Présenté par : Ahmed Yahi

Sujet

Clustering des données de puces à ADN

Soutenu le : 13 / 07 /2019

devant le jury d'examen composé de :

Nom et prénom Enseignant

Mr.Benazi Makhlouf

Université de M'sila

Président

Mr. Nasereddine AMROUNE

Université de M'sila

Encadreur

Mr.Kamel Mohamed

Université de M'sila

Examineur

Promotion : 2018 /20 19

Table des matières

Introduction Générale	01
Chapitre I : La Technologie de Microarray	
1.1 Introduction.....	02
1.2 Les bases de la biologie moléculaire.....	03
1.3 Historique de microarray.....	08
1.4 Définition de microarray.....	10
1.5 Les différents types de microarray	11
1.5.1 Les microarray ou filtre à haute densité	12
1.5.2 Les puces spottées.....	12
1.5.3 Les puces à oligonucléotides synthétisés in situ.....	12
1.6 Principe de fonctionnement de puce à ADN	13
1.6.1 Fabrication des puces à ADN.....	13
1.6.2 Préparation, étiquetage et hybridation de la cible.....	13
1.6.3 Le processus de numérisation.....	13
1.7 Les possibilités d'utilisation.....	15
1.8 Conclusion.....	15
Chapitre II : La classification automatique « clustering »	
2.1 Introduction.....	16
2.2 Définition de clustering	17
2.3 Les données traitées	17
2.4 Les trois étapes principales du clustering.....	18
2.4.1 La préparation des données.....	18
2.4.2 Implémentation d'algorithme de clustering	19
2.4.3 L'exploitation des résultats de l'algorithme.....	19
2.5 Distances et similarités.....	19
2.5.1 Les fonctions distance.....	20
2.5.2 Les fonctions similarités.....	22
2.6 Méthodes de clustering.....	22
2.6.1 Méthodes hiérarchiques.....	22
2.6.2 Méthodes de partitionnement.....	24
2.6.3 Méthode de clustering flou.....	28
2.6.4 Les algorithmes basés sur la densité.....	30
2.6.5 Les algorithmes basés sur la grille.....	31

2.7 Techniques d'évaluations du clustering	31
2.8 Caractéristiques des méthodes de clustering	33
2.9 Domaines d'applications du clustering	33
2.10 Conclusion	34
Chapitre III : Conception et Implémentation	
3.1 Introduction	36
3.2 L'Analyse en Composantes Principales	36
3.3 R Studio	40
3.4 Les résultats détaillés de clustering jeu de données Brest Cancer	43
3-5 Description d'Application	45
3-6 La conception globale	45
3.6.1 Importer et préparation des données	45
3.6.2 Normalisation des données	45
3.6.3 L'analyse en composantes principales	46
3.6.4 K-means	46
3.6.5 Validation de cluster	46
3.7 Descriptif des jeux de données réels pour le clustering	47
3.8 Résultats et discussions	47
3.9 Conclusion	49
Conclusion générale	50
Bibliographie	51

REMERCIEMENT

Tout d'abord, louange à « Allah » qui m'a guidé sur le droit chemin tout au long du travail et m'a inspiré les bons pas et les justes reflexes. Sans sa miséricorde, ce travail n'aura pas abouti.

A la fin de ce travail, nous tenons à remercier tous ceux qui ont contribué de près ou de loin à la réalisation de ce mémoire.

*A ce titre, nous remercions vivement Notre encadreur **Mr** : Nacer-Eddine Amroune pour ses conseils et son suivi durant la réalisation de notre projet.*

*.
Aussi, nous tenons à exprimer notre reconnaissance aux membres du jury*

A Et enfin un remerciement à tous nos enseignants, pour leurs contributions concrètes à travers l'accès à l'information et surtout pour le savoir et les efforts qu'ils ont fourni durant notre cursus d'étude.

Dédicace

Je dédie ce modeste travail

A celle qui est dans mon cœur, à celle qui a veillée pour notre confort et sacrifiée beaucoup pour notre réussite,

Ma chère mère, (que dieu me la garde)

A celui qui m'a toujours appris comment réfléchir avant d'agir, à celui qui m'a soutenu tout au long de ma vie scolaire, à celui qui n'a jamais épargné un effort pour mon bien,

Mon cher père (que dieu me le garde)

A mes chers frères : fouaz, fateh, abdallah et le petit elhadj.

A mes chères sœurs: ndjat, fatima zahra.

A ma petite fille Sérine

A mes chères amies qui m'ont accompagné tout au long de ce parcours universitaire.

A mes collègues et amis de classe

A tous les membres de la promotion d'master en informatique de l'université de Mohamed Boudiaf de M'sila.

Je tiens à dédier ce modeste travail à toute personne ayant contribué à son aboutissement, Et à tous ceux qui me

Connaissent...

Introduction générale

La technologie de microarray à maintenant permis de surveiller simultanément les niveaux d'expression de milliers de gènes au cours de processus biologiques importants et à travers des collections d'échantillons apparentés.

L'élucidation des modèles cachés dans les données d'expression génique offre une formidable opportunité pour une meilleure compréhension de la génomique fonctionnelle. De plus, le grand nombre de gènes et la complexité des réseaux biologiques augmentent considérablement les défis de la compréhension et de l'interprétation de la masse résultante de données, qui consiste souvent en des millions de mesures.

Un premier pas vers ce défi est l'utilisation de techniques de clustering, qui est essentiel dans le processus d'exploration de données pour révéler des structures naturelles et identifier des modèles dans les données sous-jacentes.

En raison de la grande quantité de gènes impliqués dans chaque expérience, l'analyse de cluster est une technique exploratoire très utile visant à identifier les gènes qui présentent de cluster est modèles d'expression similaires.

Chapitre I :
La technologie de
microarray

Chapitre I La Technologie de Microarray

1.1 Introduction

Ces dernières années, l'avènement de haute densité d'oligonucléotides et d'ADNc a eu un impact profond sur la recherche biologique et médicale. En effet, la nouvelle technologie permet l'acquisition de données qui se révèlent fondamentales dans de nombreux domaines des sciences biologiques, allant de la compréhension des systèmes biologiques complexes au diagnostic.

La recherche en biologie moléculaire évolue à travers le développement de technologies utilisées pour les réaliser. Il n'est pas possible de faire des recherches sur un grand nombre de gènes à l'aide de méthodes traditionnelles. DNA Microarray est une technologie de ce type qui permet aux chercheurs d'enquêter et de traiter les problèmes qui étaient autrefois considérés comme non traçables. On peut analyser l'expression de nombreux gènes en une seule réaction rapidement et de manière efficace. La technologie DNA Microarray permis à la communauté scientifique de comprendre les aspects fondamentaux qui sous-tendent la croissance et le développement de la vie, ainsi que d'explorer les causes génétiques des anomalies survenant dans le fonctionnement du corps humain.

Les Microarrays sont considérés comme une technologie révolutionnaire en biologie, facilitant l'étude quantitative de milliers de gènes simultanément à partir d'un seul échantillon de cellules.

Avant d'entrer dans l'analyse des données et d'expliquer les principes de certaines techniques, Nous devons faire une présentation des concepts simple de base de la biologie moléculaire pour permettre aux lecteurs, en particulier dans les domaines de la biologie ou des sciences naturelles non spécialisées. Et ceux qui s'intéressent à ce domaine, de se faire une idée générale et de comprendre les chapitres à venir.

1.2 Les bases de la Biologie Moléculaire

A. La cellule

La cellule représente l'unité structurelle et fonctionnelle chez les organismes vivants, la santé humaine de la santé de ses cellules et que tout défaut dans le travail de la cellule, s'il n'est pas corrigé par le corps, conduit à l'apparition de maladies qui peuvent être mortelles, telles que le cancer. Pour assurer cette cohérence, la cellule contient un ensemble de composants fonctionnant de manière ordonnée qui permet la production de matériaux nécessaires à la survie de la cellule et la réparation des erreurs éventuelles.

Les organismes vivants sont classés selon les composantes de leurs cellules en deux grandes familles: les organismes à noyau primitif (procaryotes), et le deuxième type est les vrais objets du noyau (eucaryotes). Pour que les deux types de cellules continuent à survivre, elles doivent produire des protéines et des enzymes (Un type de protéine fonctionne pour faciliter certaines réactions chimiques). Pour obtenir de l'énergie, construire des membranes, défendre la cellule et d'autres activités. [01]

Le principe de la fonction cellulaire en général est que l'activité de chaque gène conduit à la production d'une protéine qui joue plusieurs rôles dans la cellule. Ce principe est appelé dans la biologie moléculaire du principe central (Central Dogma), qui est généralement décrit comme: (Figure 1.1)

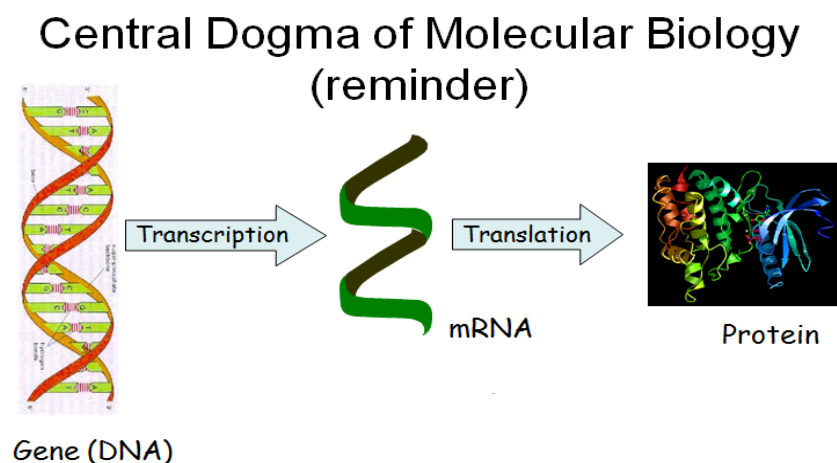


Figure 1.1 : Le principe de la fonction cellulaire (Central Dogma)

La cellule doit conserver des informations sur la façon de reconstituer cette protéine quelque part pour l'utiliser en cas de besoin, et cet endroit est un **ADN**.

B. ADN : Acide Désoxyribonucléique

L'ADN est une molécule très longue, composée d'une succession de nucléotides accrochés les uns aux autres, petites molécules simples composées d'un sucre, d'une base azotée et d'un phosphate. Il existe quatre nucléotides différents : A (adénine), C (cytosine), T (thymine) ou G (guanine), dont l'ordre d'enchaînement est très précis et correspond à l'information génétique.

L'ADN est formé de deux brins complémentaires enroulés en hélice. L'ADN joue un grand rôle dans l'hérédité il est considéré comme support de l'information génétique, il porte des milliers de gènes sous la forme de succession de nucléotides. L'information génétique n'est ainsi jamais perdue, et peut se transmettre aux nouvelles générations via les cellules germinales (Figure 1.2)

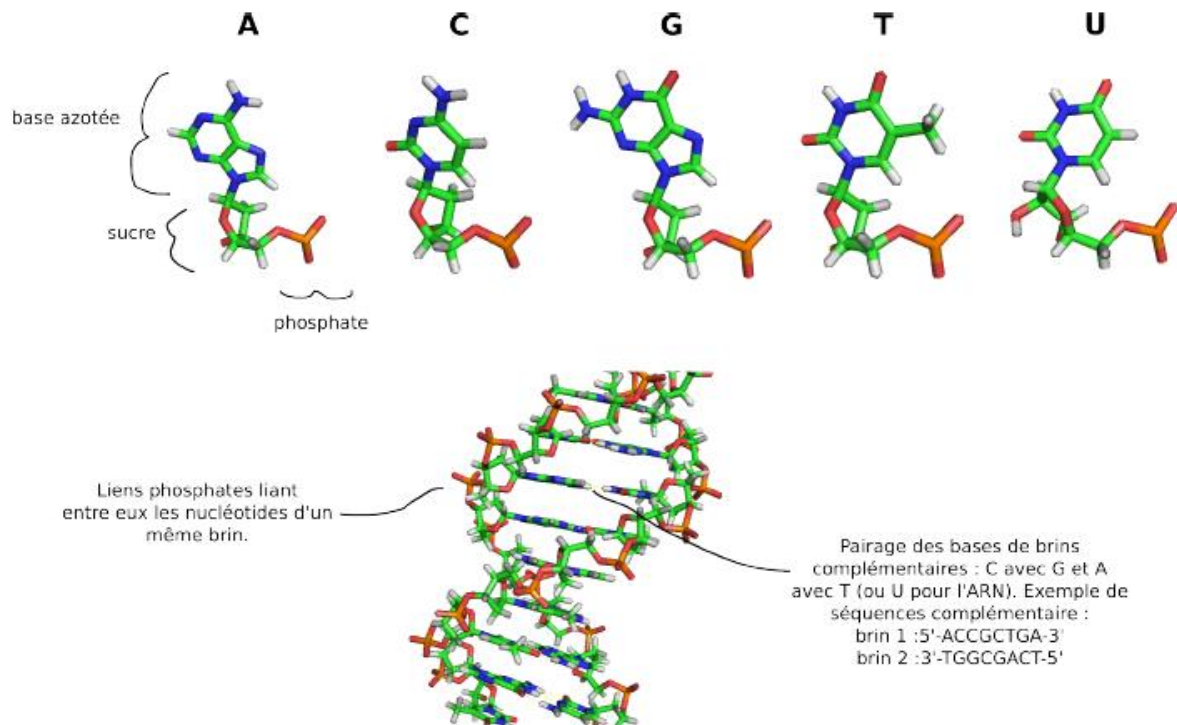


Figure 1.2 : Nucléotides et ADN sous sa forme double brin

C. ARN : Acide Ribonucléiques

L'ARN (Acide Ribonucléique) est présent chez l'ensemble des organismes et représente le produit de l'expression des gènes. L'ARN est constitué par un enchaînement d'unités élémentaires dit les ribonucléotides, est d'une suite de 4 nucléotides (A, U, C et G) ce qui permet de transférer et de traiter l'information dans la cellule le plus souvent formé d'un simple brin on distingue les ARN

messagers ou (ARNm) : ils sont transcrits à partir d'un gène (ADN). Ils sont ensuite traduits en protéines.

D. Les protéines

Les protéines sont essentielles au fonctionnement de l'organisme. Une protéine est constituée par un ou plusieurs polypeptides. Chaque polypeptide est une chaîne d'acides aminés, reliée entre elle par une liaison covalente, la liaison peptidique. Il existe vingt acides aminés différents. Dans un polypeptide, les acides aminés réagissent avec leurs voisins selon leurs affinités chimiques. La chaîne polypeptidique adopte alors une forme caractéristique qui lui confère ses propriétés. Un changement de forme s'accompagne généralement d'une perte de la fonction de la protéine, dites dénaturée.

Les fonctions des protéines sont diverses. Elles sont le reflet de l'information génétique présente dans un gène.

Les chaînes nucléotidiques possèdent deux extrémités distinctes : on peut donc les représenter de manière orientées de l'extrémité dite 5' vers l'extrémité dite 3'.

En conséquence, les chaînes nucléotidiques sont écrites sous forme d'une succession ordonnée et orientée de lettres qui représentent les unités élémentaires (les nucléotides) : (Figure 1.3)

- ADN : 4 nucléotides = C, G, A et T
- ARN : 4 nucléotides = C, G, A et U

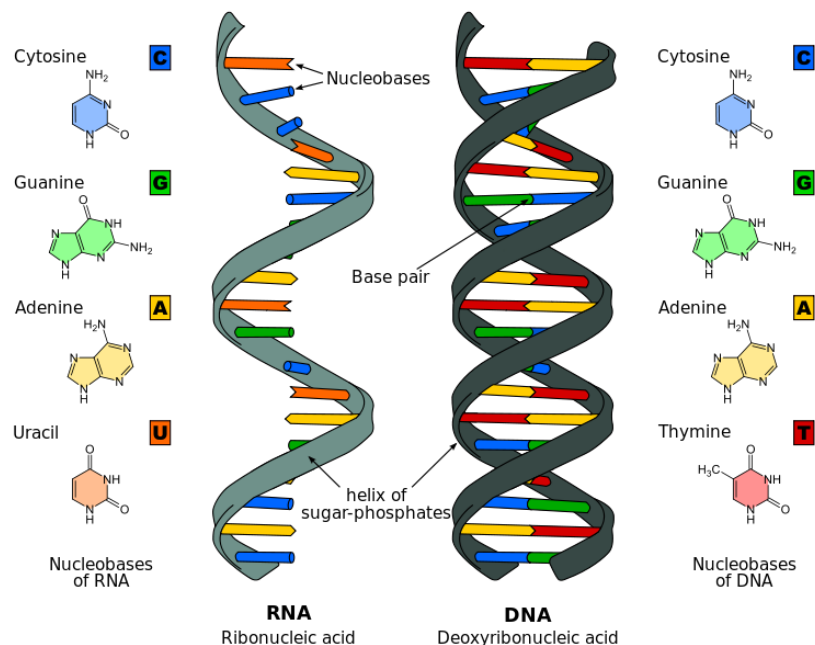


Figure 1.3 : Illustration de la structure de l'ADN et de l'ARN messager

Des informations sur la composition des protéines ont été stockées dans des parties de l'ADN appelées gène, lequel est d'abord transcrit en ARN, de sorte que l'ARN contient les molécules de base des constituants basaux du gène, mais T (Thymine) est remplacé par U (Uracile).

Après avoir formé l'ARN messenger, ce dernier se déplace hors du noyau pour être transformé en une protéine par le ribosome. Le processus de lecture du contenu d'ARN et de sa conversion en une protéine est appelé le processus de traduction.

Dans ce processus, les trois parties de base sont converties en acide aminé, le ribosome lit l'ARN et forme une chaîne d'acides aminés qui forment la protéine ou le peptide.

Le processus de transcription se déroule en deux étapes: dans la première étape, toute la séquence du gène est répliquée sous la forme d'un ARN primaire, puis une partie de cet ARN initial est enlevées afin de former l'ARN messenger à l'aide des protéines spliceosomes. Les parties capturées sont appelées introns et les parties retenues par les exons (Figure 1.4).

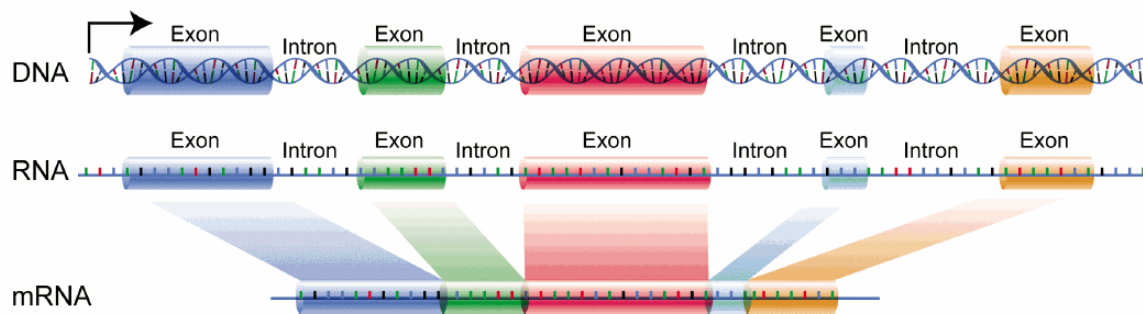


Figure 1.4 : Une illustration du principe de coupe afin de créer différentes protéines à partir d'un gène.

E. Rappel sur la transcription et la traduction

Schématiquement, les deux grandes étapes de l'ADN aux ARN messagers puis des ARN messagers aux protéines sont :

➤ La transcription

Synthèse de l'ARN messenger à partir de l'ADN. Après la transcription, l'enchaînement des 4 nucléotides de l'ARN messenger (C, G, A et U) correspond exactement à celui des 4 nucléotides (C, G, A et T) des exons de l'ADN.

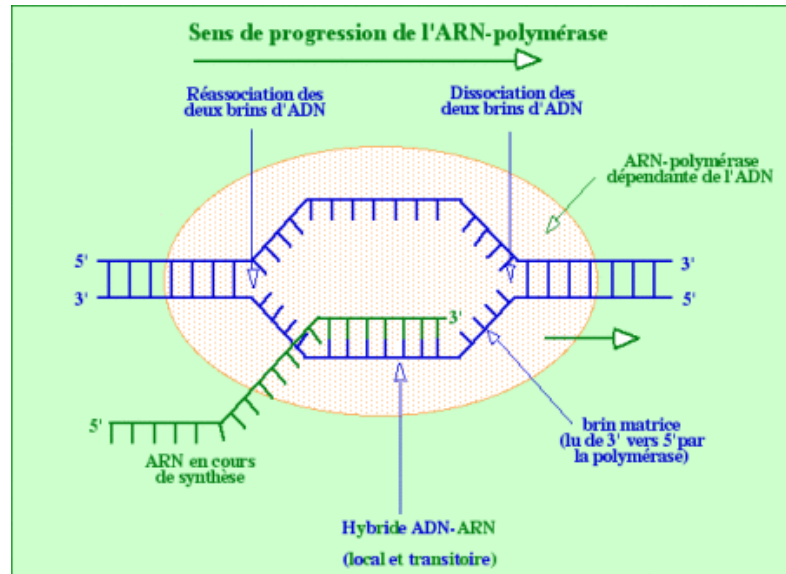


Figure 1.5 : De l'ADN à l'ARN messenger (ARNm)

➤ La traduction

Synthèse de la protéine à partir de l'ARN messenger. L'enchaînement des nucléotides de l'ARN messenger est décodé dans les ribosomes par triplet : 3 nucléotides = 1 codon. Après la traduction, l'enchaînement des 20 acides aminés de la protéine correspond exactement à celui des codons de l'ARN messenger.

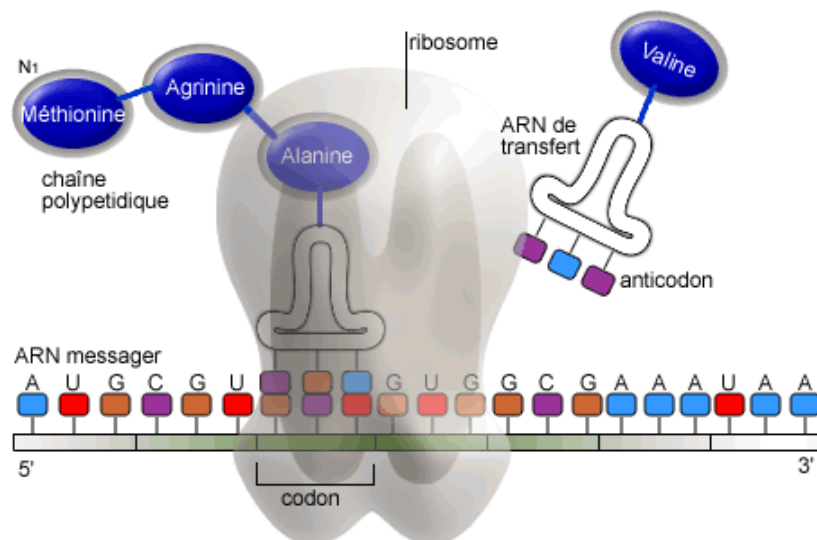


Figure 1.6 : Principe général de la traduction de l'ARNm en protéine.

Nous concluons que plus l'activité génétique d'un gène, plus la quantité d'ARNm (ARN messenger) produite est élevée et plus la proportion de protéines fabriquées par ce gène. Et de là, nous pouvons

identifier les gènes actifs dans la cellule. Le moyen le plus simple consiste à extraire l'ARNm de la cellule et en à mesurer le pourcentage, ce qui signifie si sa structure est complémentaire à celle du gène et possède une propriété importante qui peut être hybridée, nous pouvons en produire une version complémentaire (ADNc) à l'aide des molécules de base C, G, T, A.

Le phénomène de l'hybridation permet d'identifier une séquence de nucléotides, c'est-à-dire l'enchaînement des bases d'un fragment d'ADN en mettant ce dernier en présence d'autres brins d'ADN, dont la séquence est connue.

Par exemple, face à des brins d'ADN synthétique représentatifs d'une maladie, les brins extraits de l'ADN du patient vont hybrider si le malade est porteur de l'affection recherchée.

La technique microarray est basée sur ce principe hybridation afin de mesurer la proportion de l'activité des gènes d'une manière parallèle afin que, dans une expérience, nous puissions mesurer la proportion d'activité de tous les gènes dans le génome humain.

1.3 Historique de Microarray

Les premières puces à ADN sont apparues en 1993, les premiers articles dans lesquels le terme «microarray» était utilisé dans son sens actuel ont été publiés par le laboratoire de Pat Brown à l'Université de Stanford en année 1995 -1996, mais leur concept date de 1987 (cité dans Bellis et Casellas, 1997). Suite logique aux anciennes méthodes de northern blotting (Alwine et al , 1977) et d'expression différentielle (Liang et Pardee, 1992), la technologie des puces à ADN est basée sur le principe d'hybridation développé par Southern (1974). Ce principe stipule que deux fragments d'acides nucléiques complémentaires peuvent s'associer et se dissocier de façon réversible sous l'action de la chaleur et de la concentration saline du milieu. Il s'agit simplement d'une miniaturisation du système classique de reverse dot blot (Lennon et Lehrach, 1991) qui a vu le jour grâce à une technologie pluridisciplinaire intégrant l'électronique (techniques de dépôt), la chimie (préparation des lames et greffes des sondes oligonucléotidiques ou synthèse in situ), l'analyse d'images (acquisition des données) et l'informatique (interprétation des données). Depuis leur apparition, les puces à ADN suscitent un intérêt inversement proportionnel à leur taille, avec pour preuve l'explosion du nombre de publications qui leur sont dédiées L'adoption rapide de cette technique est illustrée par la publication en 2001 de plus de 900 articles sur l'utilisation de la technique des micropuces [02].

Historiquement les macroarrays, les microarrays et les « véritables » puces à ADN correspondent à trois méthodes différentes d'analyse (Lagoda et Regad, 2000). Les macroarrays utilisaient des clones d'ADN complémentaire (ADNc) disposés sur des membranes de nylon (avec un espacement de l'ordre du millimètre) en association avec des cibles radioactives. Les microarrays, plus miniaturisés, comportaient quelques milliers de gènes représentés par des produits PCR déposés tous les 200 à 400 microns sur une lame de verre et des cibles marquées par fluorescence. Enfin, les « véritables » puces à ADN associaient à chacun des gènes d'un organisme un ensemble d'oligonucléotides synthétisés in situ. La première de ces puces à ADN s'appelait la « Gene ChipTM HIV PRT ». Commercialisée en 1998 par Affymetrix (Santa Clara, CA, USA), elle avait été conçue pour l'analyse des mutations de la transcriptase inverse et de la protéase du virus HIV (cité dans Hinfrey, 1997). La même année a vu le développement de la première puce à oligonucléotides dédiée à une bactérie, contenant un sous-ensemble de cent gènes de streptococcus pneumonie (De Saizieu et al. 1998).

A titre historique et conceptuel, il est malgré tout important de connaître les principes des études du transcriptome par la technique des puces à ADN. La première puce à ADN (Figure 1.7) avec 45 sondes fluorescentes d'*Arabidopsis thaliana* est apparue en 1995 (Schena *et al.*, 1995) [03].

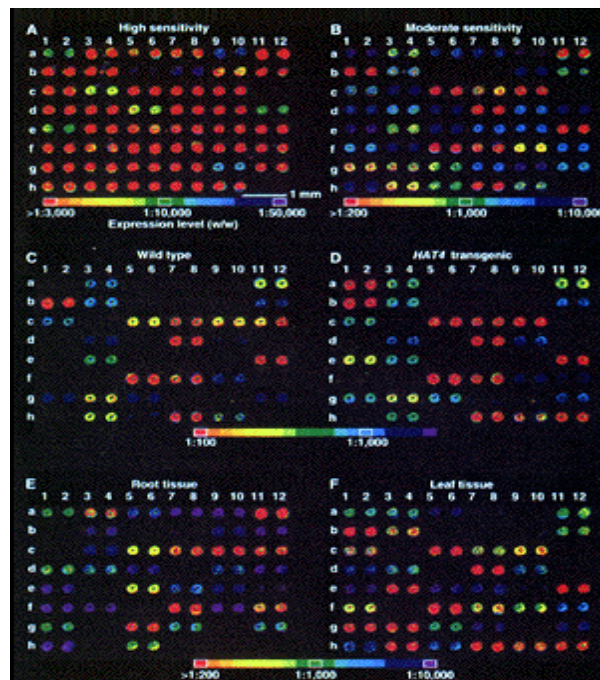


Figure 1.7 : La première puce à ADN

1.4 Définition de Microarray

Une puce à ADN est un ensemble de molécules d'ADN fixées en rangées ordonnées sur une petite surface qui peut être du verre, du silicium ou du plastique. Cette biotechnologie récente permet d'analyser le niveau d'expression des gènes (transcrits) dans une cellule, un tissu, un organe, un organisme ou encore un mélange complexe, à un moment donné et dans un état donné par rapport à un échantillon de référence.

Aujourd'hui, la puce à ADN communément appelée « DNA microarray » en anglais (de « array » = rang ordonné), désigne une technologie miniaturisée de biologie moléculaire permettant l'analyse d'échantillons biologiques complexes. Son concept repose sur un procédé multidisciplinaire intégrant la biologie, la chimie des acides nucléiques, les nanotechnologies, l'analyse d'images et la bioinformatique. Une puce à ADN est constituée de fragments d'ADN immobilisés sur un support solide appelées sondes. Ces sondes seront caractéristiques d'un gène donné ou d'une région d'ADN selon une disposition ordonnée. La position de ces fragments sur la biopuce est connue. Son fonctionnement repose sur le même principe que des technologies telles que le Southern blot ou le northern blot, qui est couramment utilisées pour détecter et quantifier la présence d'une séquence nucléique spécifique au sein d'un échantillon biologique complexe, par hybridation à une sonde de séquence complémentaire portant un marquage radioactif. La confection des puces à ADN a permis d'étendre ce principe à la détection simultanée de milliers de séquences en parallèle. Il est possible de détecter et de quantifier l'ensemble des cibles que contient ce mélange en une seule expérience.

Une puce comporte quelques centaines à plusieurs dizaines de milliers d'unités d'hybridation appelées « spots » (de l'anglais spot=tache), chacune étant constituée d'un dépôt de fragments d'ADN ou d'oligonucléotides correspondant à des sondes de séquences données. L'hybridation de la Puce avec un échantillon biologique, marqué par un radioélément ou par une molécule fluorescente, permet de détecter et de quantifier l'ensemble des cibles qu'il contient en une seule expérience.

D'abord conçues sur des membranes poreuses de nylon (appelées parfois « macroarrays » par opposition aux « microarrays »), les puces à ADN ont été progressivement mises au point sur lames de verre à la fin des années 90. La miniaturisation, rendue possible par l'utilisation d'un support solide, de marqueurs fluorescents et par les progrès de la robotique, permet aujourd'hui de fabriquer des puces comportant une très haute densité de spots, susceptibles de recouvrir l'intégralité du génome d'un organisme sur une simple lame de microscope. Parmi les principaux fabricants de puces à ADN on trouve *Affymetrix*, le leader historique des puces à ADN situé (USA - californie)

Spécialisée dans l'analyse génétique, pour 1,3 milliard de dollars et *Agilent* technologie.

1.5 Les différents types de puces à ADN

On distingue plusieurs types de puces selon la densité des spots, le mode de fabrication, la nature Des fragments fixés à la surface et les méthodes d'hybridation. Les caractéristiques des puces les plus courantes sont résumées dans **le Tableau 01**.

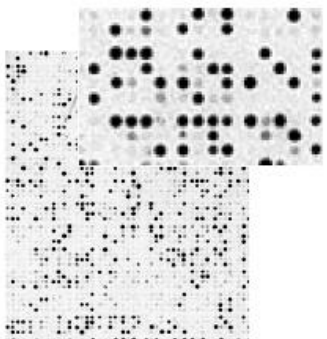
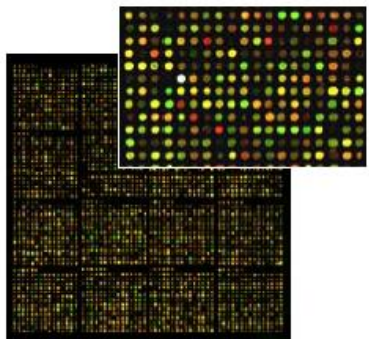
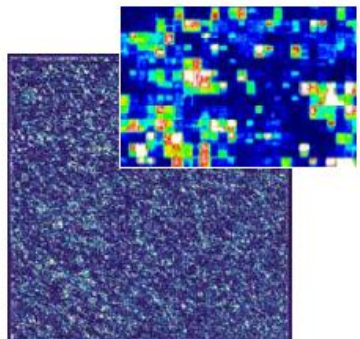
« Macroarray »	« Microarray spottée »	« GeneChips » de Affymetrix
 - support : membrane de nylon - taille des spots : 0,5-1mm - densité : quelques centaines de spots/cm² - sondes : produits de PCR - cibles : ADNc avec marquage radioactif au ³²P - principales applications : analyse de l'expression des gènes	 - support : lame de verre à revêtement chimique - taille des spots : ~100µm - densité : 1000-10000 spots/cm² - sondes : produits de PCR ou oligonucléotides longs (30-70mers) - cibles : ADNc ou produits de PCR avec marquage fluorescent au Cy3 et Cy5 - principales applications : analyse de l'expression, ChIP-on-Chip, CGH-array	 - support : lame de verre à revêtement chimique - taille des spots : ~20µm - densité : jusqu'à 250000 spots/cm² - sondes : oligonucléotides courts (20-25 mers) synthétisés <i>in situ</i> - cibles : ARNc ou produits de PCR avec marquage fluorescent à la biotine-streptavidine - principales applications : analyse de l'expression, détection de marqueurs moléculaires

Tableau 1.1 : Principaux types de puces à ADN

Les premières puces mises au point, appelé macroarray, étaient sur des membranes de nylon en 1995. Depuis le début des années 2000, elles sont réalisées sur des lames de verres. L'utilisation d'un support solide, de marqueurs fluorescents et des progrès de la robotique permettent de fabriquer des puces avec une très haute densité de spots, susceptible ou capable de couvrir l'ensemble du transcriptome d'un organisme sur une lame de microscope. Le revêtement le plus couramment utilisé est la polylysine, qui permet l'adhésion de l'ADN.

1.5.1 Les microarray ou filtre à haute densité

Les dépôts (sondes) sont des clones d'ADNc ou des produits de PCR fixés sur une membrane de nylon de 8 par 12 cm. Le marquage des cibles est radioactif (P32). Elles permettent une mesure de l'abondance relative des ARNm présents dans l'échantillon de départ.

Les filtres à haute densité (macroarrays) sont des plaques en nylon (12X8 cm) qui permettent de quantifier la présence de 2400 gènes par marquage radioactif d'un échantillon (ARNm totaux à étudier).

1.5.2 Les puces spottedées

Les sondes sont des ADN doubles brin de 200 à 2000 paires de bases, amplifiées par PCR ou plus récemment des oligonucléotides longs (50 à 70 mères (unités)). Des microgouttelettes (spots) de ces solutions de sondes sont déposées selon une matrice d'emplacement défini à une densité supérieure ou égale à 1000 sondes/cm² par un robot sur des lames de verre (jusqu'à 12000 sondes). En général, un spot correspond à un gène donné.

Dans le cas de produit de PCR, l'ADN de la puce doit être dénaturé pour obtenir de l'ADN simple brin (une transcription inverse).

1.5.3 Les puces à oligonucléotides synthétisés in situ (Gene chips)

Ces puces (de 1,28cm x 1,28cm) renferment des oligonucléotides simple brin synthétisés in situ (directement sur le support) par photolithographie. Ce type de puce permet donc une miniaturisation à l'extrême (250 000 spots/cm²) et une grande précision dans l'analyse du gène est représentée, car une quinzaine de sonde couvre un même gène, mais sur des portions différentes. [04]

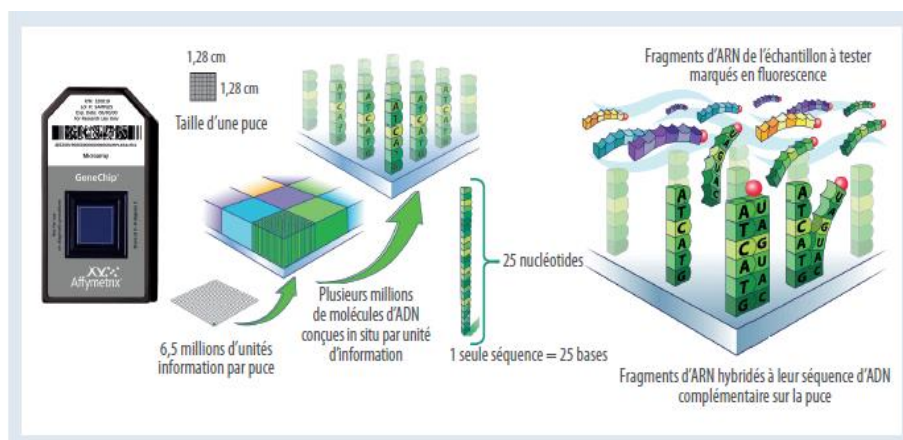


Figure 1.8 : Les puces à oligonucléotides synthétisés in situ (« Gene chips » Affymetrix).

1.6 Principe de fonctionnement de Puce ADN

Par rapport à l'approche traditionnelle de la recherche en génomique, qui a mis l'accent sur l'examen local et la collecte de données sur des gènes uniques, les technologies de la microarray ont maintenant permis de surveiller les niveaux d'expression de dizaines de milliers de gènes en parallèle. Les deux principaux types de microarray les expériences sont le microarray d'ADNc et les matrices d'oligonucléotide (oligo puce) . Malgré les différences dans les détails de leurs protocoles d'expérimentation, les deux types d'expériences impliquent trois procédures de base communes :

1.6.1 Fabrication des puces à ADN : un microarray est une petite puce (faite de verre enduit chimiquement, de nylon membrane ou silicium), sur laquelle sont attachés des dizaines de milliers de molécules d'ADN (sondes) dans les grilles fixes. Chaque cellule de grille se rapporte à une séquence d'ADN.

1.6.2 Préparation, étiquetage et hybridation de la cible : typiquement, deux échantillons d'ARNm (un échantillon d'essai et un échantillon de contrôle) sont transcrits à l'envers en ADNc (cibles), étiquetés en utilisant soit colorants ou isotopiques radioactifs, puis hybridé avec les sondes à la surface de la Puce.

1.6.3 Le processus de numérisation : les puces sont scannées pour lire l'intensité du signal émise par cibles marquées et hybridés.

Les puces à ADN permettent de mesurer et de visualiser très rapidement les différences d'expression entre les gènes et ceci à l'échelle d'un génome complet. Si la mise en œuvre de la technique est assez compliquée, son principe est très simple. En voici les principales étapes:

- Fabrication de la puce
- Préparation de la cible
- L'hybridation
- La lecture
- L'analyse des données
- Le regroupement des profils d'expression

La mesure d'expression des gènes procède en trois étapes. En premier lieu, la fabrication de la puce consiste à déposer des sondes sur la lame (figure de gauche). La deuxième étape est la technique d'hybridation, qui va mettre les cibles au contact des sondes. Dans l'exemple (figure du milieu), la technique utilise deux canaux (marquage par les fluorochromes Cy3 et Cy5). La dernière étape est

l'acquisition par scanner d'une image de la puce. Cette image est analysée pour détecter les spots et quantifier les niveaux d'expressions dans les deux canaux.

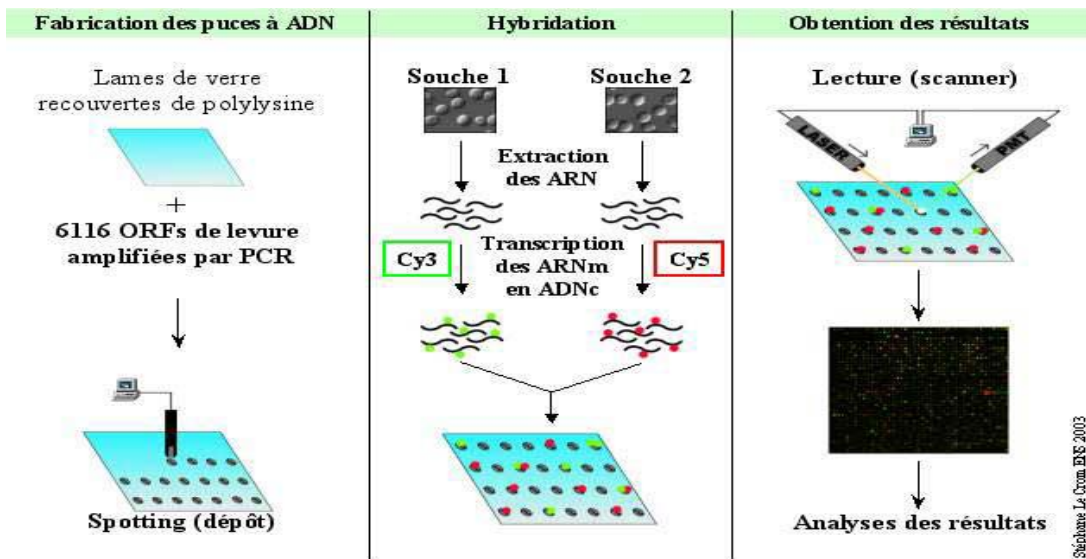


Figure 1.9 : Principe des puces à ADN

1.7 Les possibilités d'utilisation

Les puces à ADN permettent des tests plus rapides, plus sensibles et plus spécifiques. En évitant certaines étapes préliminaires telle que la culture, cela permet d'obtenir un résultat en quelques heures là où plusieurs jours étaient nécessaires.

➤ L'analyse de mutations

A côté de la séquence sauvage (sonde de référence), on utilise plusieurs sondes identiques à une mutation près. Ces sondes vont couvrir l'ensemble des mutations potentielles (substitution, délétion, addition). De manière générale, l'analyse de substitutions (pour une position donnée) implique donc l'utilisation de quatre sondes, une pour la séquence sauvage et trois pour chacun des nucléotides pouvant se substituer au nucléotide original. Cela permet ainsi de repérer des séquences mutantes dans un échantillon test.

➤ Le séquençage par hybridation

L'analyse procède par lecture de petits blocs. L'analyse porte sur des sous-séquences chevauchantes qui sont lues et réassemblées au moyen d'un programme informatique qui reconstitue la séquence étudiée.

➤ Le diagnostic des maladies génétiques

Les immuno-essais (hormone et cancer), en microbiologie (antigènes/anticorps), les facteurs aggravants des maladies cardio-vasculaires et du vieillissement. Etude d'oncogènes spécifiques.

➤ **Le diagnostic des maladies infectieuses**

Détection de microorganismes (présence, mutations, capacité à résister à certains antibiotiques ou à des inhibiteurs d'enzymes).

➤ **La pharmacogénomique**

Identification de cibles pour la recherche thérapeutique.

➤ **La toxicogénomique**

Dans le cas de l'analyse de la réponse toxicologique à l'aide de puces à ADN, on doit au départ faire l'hypothèse que l'effet toxicologique, direct ou indirect, d'une substance résulte de la modification de l'expression de certains gènes, ou du moins qu'il lui est associé. Les mesures consistent alors à évaluer l'expression différentielle de gènes de deux échantillons, grâce à un marquage par fluorescence en deux couleurs (une par échantillon). Cela permet de comparer les profils d'expression génique de deux tissus différents (normal/pathologique ou traité/non traité par une substance).

➤ **L'agro-alimentaire**

Contrôle des microorganismes utilisés dans certaines fabrication (ferments lactiques, levures, mycélium, etc...), détection des séquences provenant d'organismes génétiquement modifiés (OGM) dans les semences ou dans certains aliments, détection d'agents infectieux dans l'alimentation comme la Salmonelle ou la Listeria.

➤ **L'environnement**

Analyse bactérienne de l'eau de consommation, détection des agents infectieux dans l'alimentation, l'air ou l'eau (Salmonella, Listeria, Legionella).

2.8 Conclusion

Le profil d'expression génique à base de microarray d'ADN repose sur l'hybridation de l'acide nucléique et sur l'utilisation de l'acide nucléique immobilisé sur une surface solide. Des techniques de profilage d'expression ont été utilisées pour surveiller simultanément l'expression de milliers de gènes à partir d'échantillons biologiques. Ils sont relativement faciles à utiliser et peuvent être appliqués à un grand nombre d'échantillons en parallèle. Divers outils de calcul ont été créés pour faciliter l'analyse du volume important de données produites dans des expériences de microarray d'ADN. La normalisation est une étape cruciale pour l'obtention de données fiables et utilisables pour une analyse ultérieure, comme l'identification de gènes et de clusters exprimés de façon différenciable. Nous décrivons ici, les différentes technologies de microarray, leur utilisation et les différentes manières d'exprimer les données en tant que modèles d'expression de gènes.

Chapitre II :
La classification
automatique «
clustering »

2.1 Introduction

En marketing, on va chercher à regrouper des personnes ayant des comportements de consommation similaires pour cibler par exemple une campagne publicitaire ; En biologie, on cherche à identifier des groupes de gènes ayant le même comportement pour mettre en place des thérapies géniques.....etc.

Dans de nombreux domaines d'application de l'analyse de données ou de la recherche d'information, il est utile de grouper de façon non supervisée des objets par similarité sans qu'il soit aisé de les représenter par des vecteurs de propriétés numériques. En biologie moléculaire, la similarité permet de capturer soit la structure complexe des objets (cas des molécules ou des séquences) soit la sémantique de leur description (cas des maladies ou des gènes). De nombreuses mesures de similarité ont été proposées mais une fois converties en dissimilarité en vue du clustering, ces mesures ne présentent pas forcément les bonnes propriétés d'une métrique. Le clustering d'objets pour lesquels on ne dispose que d'une matrice de dissimilarité requiert d'utiliser des méthodes adéquates. Nous proposons ici une évaluation comparative du clustering de gènes sur la base d'une mesure de similarité de la gène.

Le clustering c'est la tâche qui consiste à regrouper des éléments de manière automatique. Cette approche consiste à faire émerger des groupes au sein d'un ensemble d'éléments sans aucune information a priori dans ce cas, cette tâche est appelée, selon les domaines, classification non supervisée, classification automatique ou encore en utilisant l'anglicisme clustering. Les groupes créés sont appelés clusters.

Ce problème a été abordé dans de nombreux contextes et par des chercheurs dans beaucoup de disciplines, ce qui reflète son attrait et son utilité comme l'une des étapes les plus importantes de l'analyse exploratoire des données.

Dans ce chapitre, nous nous intéressons qu'aux techniques de classification automatique (clustering) et nous montrons, quels sont leurs avantages et difficultés. En tentant décrire quelques remèdes et de présenter l'avantage de Clustering dans le domaine de traitement de données non-étiquetées (sans connaissance préalable).

2.2 Définition du clustering

Le clustering est la tâche qui consiste à regrouper un ensemble d'objets (physiques ou abstraits) ou plus largement de données, de façon non supervisée (c.-à-d. sans l'aide préalable d'un expert), de telle manière que les objets d'un même groupe (appelé cluster) sont plus proches (au sens d'un critère de (dis)similarité choisi) les uns aux autres que celles des autres groupes (clusters).

Il s'agit d'une tâche principale dans la fouille exploratoire de données, et une technique d'analyse statistique des données très utilisée dans de nombreux domaines, y compris l'apprentissage automatique, la reconnaissance de formes, le traitement de signal et d'images, la recherche d'information, la bio-informatique, la compression de données et l'infographie. Etc.

L'idée est donc de découvrir des groupes au sein des données, de façon automatique. [05]

Pour établir l'équilibre, il minimise l'inertie à l'intérieur des classes et maximise celle entre les sous-groupes afin de bien les différencier. L'objectif peut être de hiérarchiser ou de répartir les données.

La qualité d'un résultat de clustering dépend de la mesure de similarité utilisée par la méthode et ses la mise en œuvre.

Le clustering peut également être utilisé pour la détection des valeurs aberrantes, où les points aberrants (qui sont «éloignés» de tout cluster) peuvent être plus intéressants que les cas courants. Par exemple la détection de la fraude par carte de crédit et la surveillance des activités criminelles dans le commerce électronique.

Pour être appliqué, le clustering s'appuie sur des algorithmes plus ou moins complexes, tels que les algorithmes des k-moyennes ou k-medoids, ou les algorithmes de maximisation de l'espérance.

2.3 Les données traitées

Les données traitées en classification automatique peuvent être des images, signaux, textes, autres types de mesures, etc. Dans le cadre de ce travail les données seront des données multidimensionnelles, par exemple une image (couleur). Chaque donnée est donc composée de plusieurs variables (descripteurs). Pour le cas de données multidimensionnelles standard, chaque donnée étant de dimension d (donc un point dans l'espace $\mathbb{R}^d$) et éventuellement étiquetée (dans le cas où l'on connaîtrait sa classe d'appartenance (le "label")) et peut donc être modélisée, par exemple, par une structure contenant au moins les coordonnées du point et le champ "label".

Les n données peuvent donc être modélisées comme étant un tableau de n éléments, chaque élément du tableau étant une structure "donnée" comme décrite précédemment. Dans le cas d'une image (couleur), l'image contenant n lignes et m colonnes et donc $(n \times m)$ pixels couleurs, chaque pixel est

composé de trois composantes RVB (Rouge, vert, bleu), peut être modélisée par un tableau de $(n \times m)$ structures. Chaque structure "pixel" est composée au moins des champs, couleur et "label".

2.4 Les trois étapes principales du clustering

Le processus de clustering se divise en trois étapes majeures :

- (1) **La préparation des données.**
- (2) **L'algorithme de clustering.**
- (3) **L'exploitation des résultats de l'algorithme.**

Nous discutons ici des problématiques générales liées à chacune de ces étapes. [06]

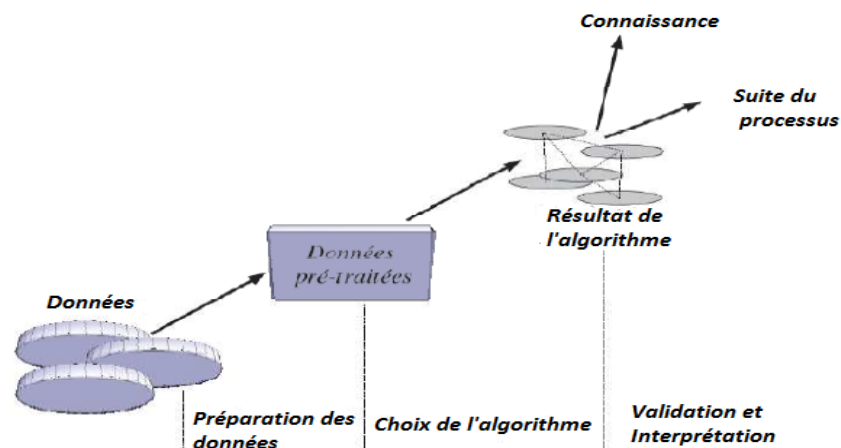


Figure 2.1 : Les différentes étapes du clustering.

2.4.1 La préparation des données

Les objets sont décrits par des variables. Ces variables sont de différentes natures :

- variables quantitatives : continues (e.g. la taille d'une personne), discrètes (e.g. le nombre de personnes) ou sous forme d'intervalles (e.g. la période de vie d'une personne).
- variables qualitatives : non-ordonnées (e.g. la "couleur" des cheveux) ou ordonnées (e.g. la taille : "petit", "moyen", "grand", etc.).
- variables structurées : par exemple la forme d'un objet (polygone, parallélogramme, rectangle, ovale, cercle, etc.)

L'étape de la préparation des données consiste à sélectionner et/ou pondérer ces variables, voire à créer de nouvelles variables, afin de mieux discriminer entre eux les objets à traiter non-supervisé. [08]

2.4.2 Implémentation d'algorithme de clustering

Le choix de l'algorithme de clustering doit donner lieu à une analyse globale du problème : quelle est la nature (qualitative et quantitative) des données ? Quelle est la nature des clusters attendus (nombre, forme, densité, etc.) ? L'algorithme doit alors être choisi de manière à ce que ses caractéristiques répondent convenablement à ces deux dernières questions.

Les critères de décision peuvent être : la quantité de données à traiter, la nature de ces données, la forme des clusters souhaités ou encore le type de schéma attendu (pseudo-partition, partition stricte, dendrogramme, etc.)

2.4.3 L'exploitation des résultats de l'algorithme

Les clusters obtenus ne sont généralement ni remis en cause, ni évalués en terme de disposition relative, dispersion, orientation, séparation, densité ou stabilité. Pourtant, il est sans aucun doute utile de distinguer les classes pertinentes obtenues, des autres. De même, cette étape d'analyse permet d'envisager le recours à une autre approche de clustering plus adaptée. Deux situations sont possibles : soit la tâche de clustering s'inscrit dans un traitement global d'apprentissage, soit les clusters générés par clustering constituent un résultat final.

Dans le premier cas, l'analyse des clusters obtenus (mesures statistiques de qualité) peut aider à orienter le traitement suivant. Une description des clusters n'est pas nécessaire dans cette situation.

En revanche, dans le cas où le clustering constitue à lui seul un processus global de découverte de classes, l'exploitation des clusters pour une application donnée passe par une description de ces derniers. Lorsque les objets se présentent sous la forme d'une matrice de (dis)similarité, il existe peu de méthodes pour décrire les classes (medoides, k objets représentatifs et mesures de cohésion). Lorsque les objets sont décrits par un ensemble de variables, on peut avoir recours à des méthodes de description des classes (descriptions conceptuelles).

2.5 Distances et similarités

La plupart des algorithmes de clustering utilisent une mesure de proximité entre les objets à traiter. Cette notion de "proximité" est formalisée à l'aide d'une mesure (ou indice) de similarité, dis-similarité ou encore par une distance. Le choix ou la construction de cette mesure est déterminant pour la suite du processus. En effet, deux mesures différentes peuvent induire deux schémas de classification différents. Finalement, chaque domaine d'application possédant ses propres données, il possède également sa propre notion de "proximité" ; il faut concevoir alors une mesure différente pour chaque domaine d'application,

permettant de retranscrire au mieux les différences (entre les objets) qui semblent importantes pour un problème donné. [08]

Donc, chaque méthode de clustering est basée uniquement sur une mesure de distance ou similarité

- Que se passe-t-il si les données ne sont pas proprement transformées ?
- Que se passe-t-il s'il y a des points extrêmes ?
- Que se passe-t-il avec des effets de saturation ?

Même avec un grand nombre d'échantillons, une mauvaise mesure de distance ou similarité ne sera pas d'un grand secours.

On distingue deux types principaux de mesures employés pour estimer cette relation [09] :

- Mesures de distance.
- Mesures de similitude.

Plusieurs méthodes de clustering utilisent les mesures de distance pour statuer sur la similitude ou la dissimilitude de n'importe quelle paire d'objets. La mesure de distance entre deux objets x_i et x_j est notée $d(x_i, x_j)$. Une telle mesure devrait être symétrique et atteindre sa valeur minimale (habituellement zéro) en cas de vecteurs identiques.

La mesure de distance est appelée mesure métrique de distance si elle aussi satisfait les propriétés suivantes:

1. Inégalité de triangle : $d(x_i, x_k) \leq d(x_i, x_j) + d(x_j, x_k) \quad \forall x_i, x_j, x_k \in S$.
2. $d(x_i, x_j) = 0 \Rightarrow x_i = x_j \quad \forall x_i, x_j \in S$.

Où S représente l'ensemble des objets.

2.5.1 Les fonctions distance

La distance $d(x_i, x_j)$ (notée $d(i, j)$) entre les points x_i et x_j , est ensuite mesurée dans cet espace grâce à l'utilisation d'une fonction de distance $d : R_M \times R_M \rightarrow R^+$, respectant les propriétés suivantes :

1. $d(i, j) \geq 0$ (contrainte positive)
2. $d(i, j) = 0$ si et seulement si $x_i = x_j$
3. $\forall x_i, x_j, d(j, i)$ (contrainte de symétrie)
4. $x_i, x_j, x_l, d(i, l) \leq d(i, j) + d(j, l)$ (contrainte d'inégalité triangulaire)

Dans la littérature, il est aisé de trouver plusieurs définitions de métrique de distance.

Cependant, toutes ces mesures ont une interprétation identique.

Dans ce qui suit, nous allons passer en revue les principales mesures de distance utilisées en clustering.

1- La distance Euclidienne :

est la plus connue et la plus utilisée. Elle peut être vue comme un cas particulier de la distance de Minkowski pour $q=2$:

$$d(i, j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2)}$$

Propriétés:

- $d(i, j) \geq 0$
- $d(i, i) = 0$
- $d(i, j) = d(j, i)$
- $d(i, j) \leq d(i, k) + d(k, j)$

2- La distance de Minkowski :

La distance de Minkowski est utilisée lorsque les attributs des objets sont de type numérique.

$$d(i, j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \dots + |x_{ip} - x_{jp}|^q)}$$

Avec q un entier positive non nul

Où $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ et $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ sont deux objets p -dimensionnels.

- Pour $q=2$, on obtient la distance Euclidienne.
- Pour $q=1$, la mesure est appelée distance de Manhattan.
- Pour $q = \infty$, c'est la métrique de Tchebychev

La mesure de distance de Minkowski peut être facilement calculée pour des attributs à Évaluation continue. Dans le cas d'objets décrits par des attributs catégoriques, binaires, ordinaux ou du type mixte, une mesure de distance spécifique doit être définie.

3- La distance de Manhattan : (appelée aussi taxi-distance)

Si $q = 1$, d est la distance de Manhattan.

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

2.5.2 Les fonctions similarités

La notion de similarité entre points, qui est plus générale. De plus, contrairement à la mesure de distance qui prend des valeurs entre 0 et l'infini, la mesure de similarité permet de fournir des valeurs quantifiables entre 0 et 1.

La fonction de similarité $s(x_i; x_j)$ qui compare deux vecteurs x_i et x_j constitue une alternative aux mesures de distance.

2.6 Méthodes de clustering :

Dans cette section, nous décrivons les algorithmes de clustering les plus connus. La raison principale pour laquelle de nombreuses méthodes de clustering sont utilisées est le fait que la notion de «cluster» n'est pas définie avec précision (Estivill-Castro, 2000). Par conséquent, de nombreuses méthodes de classification ont été développées, chacune d'elles utilisant un principe d'induction différent. (Farley et Raftery ,1998) suggèrent de diviser les méthodes de regroupement en deux groupes principaux: les méthodes hiérarchiques et les méthodes de partitionnement.(Han et Kamber ,2001) suggèrent de classer les méthodes en trois catégories principales: les méthodes basées sur la densité, les méthodes de clustering et les méthodes une catégorisation alternative basée sur le principe d'induction des différentes méthodes de regroupement est présentée dans (Estivill-Castro, 2000).

Les méthodes de clustering sont généralement classifiées en quatre catégories majeures [10] :

- Les méthodes hiérarchiques.
- Les méthodes de partitionnement.
- Les méthodes basées sur la densité.
- Les méthodes basées sur la grille.

2.6.1 Méthodes hiérarchiques

Ces méthodes construisent les clusters en partitionnant de manière récursive les instances de manière descendante ou ascendante. Ces méthodes peuvent être subdivisées comme suit:

A/ le clustering hiérarchique : d'agglomération («bottom-up»)

Dans le premier cas, on décompose l'ensemble d'individus en une arborescence de groupes.

Algorithme :

1. Initialement, mettre chaque objet dans son propre cluster ;
2. Parmi tous les clusters courants, sélectionner les deux clusters ayant la plus petite distance ;
3. Remplacer ces deux clusters par un nouveau cluster, forme par la fusion des deux clusters originaux ;

4. Répéter les étapes 2 et 3 jusqu'à atteindre la condition d'arrêt.

B/ le clustering non-hiérarchique : de division («top-down»)

Dans le deuxième, on décompose l'ensemble d'individus en K groupes, les algorithmes de ce type peuvent aussi être utilisés comme algorithmes de division dans le clustering hiérarchique.

Algorithme :

1. Mettre tous les objets dans un seul cluster ;
2. Répéter jusqu'à atteindre la condition d'arrêt :
 - (a) Choisir un cluster à diviser ;
 - (b) Remplacer le cluster choisi par le sous cluster obtenu.

Le résultat des méthodes de clustering hiérarchiques est un dendrogramme, représentant le regroupement imbriqué d'objets et les niveaux de similarité auxquels les regroupements changent. Un regroupement des objets de données est obtenu en coupant le dendrogramme au niveau de similarité souhaité.

La qualité d'une méthode de clustering hiérarchique pure souffre de son incapacité à effectuer un ajustement une fois qu'une décision de fusion ou de scission a été exécutée. En d'autres cas, si une décision de fusion ou scission particulière s'avère être un mauvais choix, la méthode ne peut pas être revenir en arrière et le corriger.

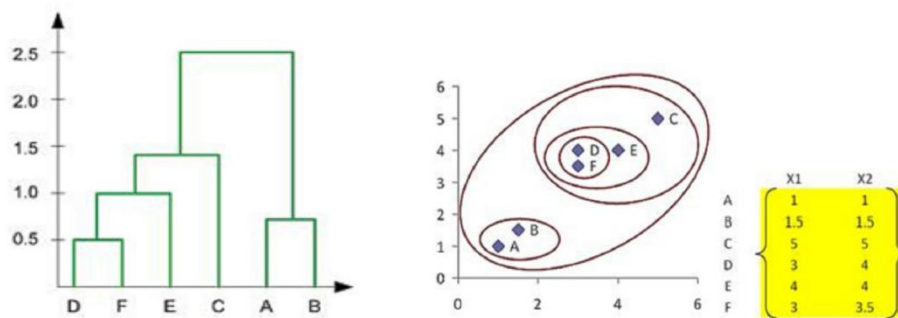


Figure 2.2 : Exemple du clustering hiérarchique

Avantages du clustering hiérarchique :

- Flexibilité incluse concernant le niveau de la granularité.
- Facilite de manipuler toutes formes de similitude ou de distance.
- Applicabilité a tout type d'attribut.

Inconvénients :

- Imprécision sur les critères d'arrête.
- La plupart des algorithmes hiérarchique ne revisitent pas les clusters une fois construits en vue de l'amélioration des résultats.
-

2.6.2 Les méthodes de partitionnement

Les méthodes de partitionnement relocalisent les instances en les déplaçant d'un cluster à un autre, à partir d'un partitionnement initial. Ces méthodes exigent généralement que le nombre de clusters sera prédéfini par l'utilisateur. Pour obtenir l'optimalité globale dans le clustering basé sur partitionné, un processus d'énumération exhaustif de toutes les partitions possibles est nécessaire. Parce que cela n'est pas faisable, certaines heuristiques gourmandes sont utilisées sous forme d'optimisation itérative. À savoir, une méthode de relocalisation déplace de manière itérative des points entre les clusters k . Les sous-sections suivantes présentent différents types de méthodes de partitionnement.

a) Principe méthode de partitionnement

Avec une base de données de n objets ou n -uplets de données, une méthode de partitionnement construit k partitions des données, où chaque partition représente un cluster, où chaque groupe doit contenir au moins un objet et chaque objet doit appartenir à un groupe exactement.

Pour obtenir une optimalité globale dans le clustering basée sur le partitionnement, il nécessiterait énumérer de manière exhaustive toutes les partitions possibles.

Il existe plusieurs méthodes de clustering par partitionnement, parmi elles on cite :

(1) l'**algorithme k-means**, où chaque cluster est représenté par la valeur moyenne des objets du cluster.

(2) l'**algorithme k-medoids**, où chaque cluster est représenté par l'un des objets situés près du centre du cluster.

Ces méthodes de clustering heuristique fonctionnent bien pour rechercher des clusters de forme sphérique. Pour trouver des clusters avec des formes complexes et pour regrouper des ensembles de données très volumineux, les méthodes basées sur le partitionnement doivent être étendues.

b) K-means

Le clustering K-means (MacQueen 1967) est l'un des algorithmes d'apprentissage automatique non supervisés les plus couramment utilisés pour la partition d'un ensemble de données ; donné en un ensemble de k clusters), où k représente le nombre de groupes prédéfinis par l'analyste. Il classe les objets en plusieurs clusters, de sorte que les objets au sein d'un même cluster ont une similitude intra-classe élevée, tandis que les objets de différents clusters sont aussi différents que possible (faible inter-classe similitude). Dans le cluster k-means, chaque cluster est représenté par son centre (centroïde) qui correspond à la moyenne des points attribués au cluster. Chaque cluster est représenté par son centre de gravité.

c) Fonctionnement de l'algorithme k-means

C'est l'un des algorithmes de clustering les plus répandus. Il permet d'analyser un jeu de données caractérisées par un ensemble de descripteurs, afin de regrouper les données "similaires" en groupes (ou clusters).

La similarité entre deux données peut être inférée grâce à la "distance" séparant leurs descripteurs ; ainsi deux données très similaires sont deux données dont les descripteurs sont très proches. Cette définition permet de formuler le problème de partitionnement des données comme la recherche de K "données prototypes", autour desquelles peuvent être regroupées les autres données.

Ces données prototypes sont appelés *centroïdes* ; en pratique l'algorithme associe chaque donnée à son centroïde le plus proche, afin de créer des *clusters*. D'autre part, les moyennes des descripteurs des données d'un cluster, définissent la position de leur centroïde dans l'espace des descripteurs : ceci est à l'origine du nom de cet algorithme (K-moyennes ou *K-means* en anglais).

Après avoir initialisé ses centroïdes en prenant des données au hasard dans le jeu de données, *K-means* alterne plusieurs fois ces deux étapes pour optimiser les centroïdes et leurs groupes :

1. Regrouper chaque objet autour du centroïde le plus proche.
2. Remplacer chaque centroïde selon la moyenne des descripteurs de son groupe.

Après quelques itérations, l'algorithme trouve un découpage stable du jeu de données : on dit que l'algorithme a convergé.

L'algorithme K-Means est répété plusieurs fois pour obtenir une solution de clustering optimale, commençant à chaque fois par un ensemble aléatoire de clusters initiaux.

Convergence ?

Ici, on dit que l'algorithme converge quand plus rien ne bouge d'une itération à l'autre, c'est-à-dire quand le centroïde reste immobile même après avoir recalculé pour chaque point son cluster.

Rien de tel qu'un petit schéma pour illustrer tout cela !

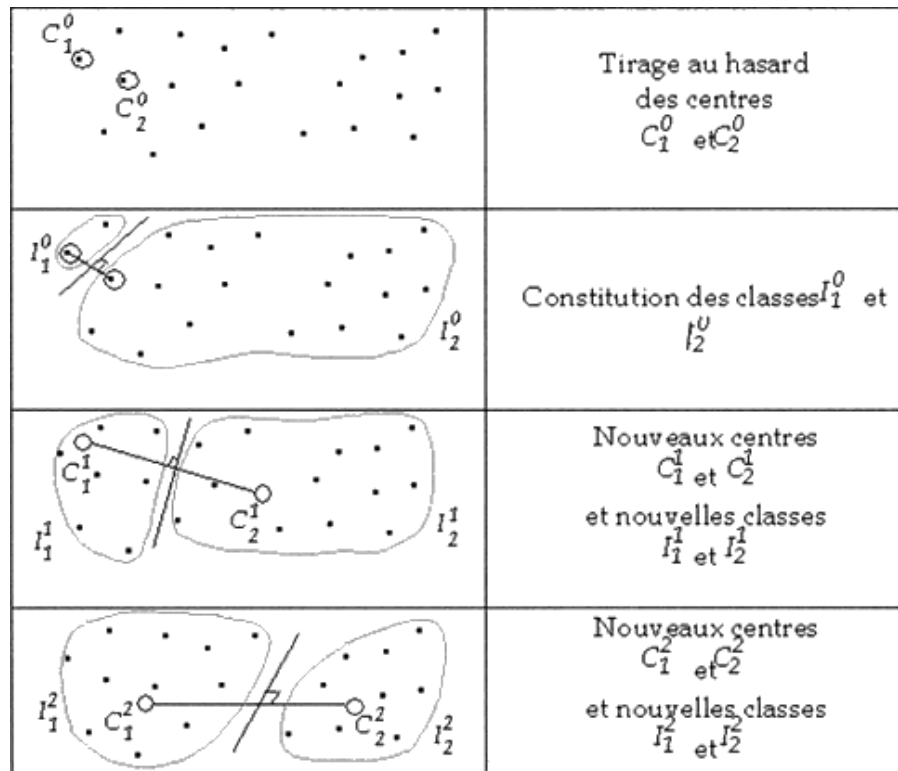
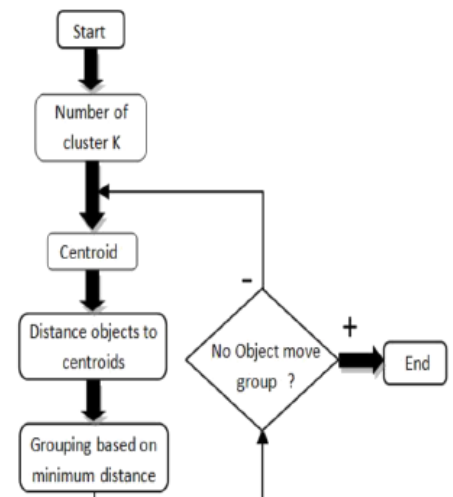


Figure 2.3 : Exemple sur l'algorithme K-mean

Comme tout algorithme, *K-means* présente des avantages et des inconvénients : il est simple, rapide et facile à comprendre ; cependant il ne permet pas de trouver des groupes ayant des formes complexes.

d) Algorithme K-means

- Choisir K éléments initiaux "centres" des K groupes
- Placer les objets dans le groupe de centre le plus proche
- Recalculer le centre de gravité de chaque groupe
- Itérer l'algorithme jusqu'à ce que les objets ne changent plus de groupe.



e) Cas d'utilisation K-means

K-Means en particulier et les algorithmes de clustering de façon générale ont tous un objectif commun : Regrouper des éléments similaires dans des clusters. Ces éléments peuvent être tous et n'importe quoi, du moment qu'ils sont encodés dans une matrice de données.

Les champs d'application de K-Means sont nombreux, il est notamment utilisé en :

- la segmentation de la clientèle en fonction d'un certain critère (démographique, habitude d'achat etc....)
- Utilisation du clustering en Data Mining lors de l'exploration de données pour déceler des individus similaires. Généralement, une fois ces populations détectées, d'autres techniques peuvent être employées en fonction du besoin.
- Clustering de documents (regroupement de documents en fonction de leurs contenus. Pensez à comment Google Actualités regroupe des documents par thématiques.)

➤ Avantages de l'algorithme k-means

- La méthode résolve une tâche non supervisée, donc elle ne nécessite aucune information sur les données.
- Sa simplicité conceptuelle et sa rapidité
- L'algorithme de k-means est très populaire du fait qu'il est très facile à comprendre et à mettre en oeuvre.
- La méthode est applicable à tout type de données (mêmes textuelles), en choisissant une bonne notion de distance.

➤ Inconvénients de l'algorithme k-means

- La difficulté de trouver une bonne fonction de distance.
- Un bon choix du nombre k est nécessaire, un mauvais choix de k produit de mauvais résultats.
- La difficulté d'expliquer certains clusters (i.e. attribuer une signification aux groupes constitués).
- Il n'est pas applicable en présence d'attributs qui ne sont pas du type numérique
- Ne peut découvrir les groupes non-convexes
- Sensible à la présence de bruits.

f) Les méthodes K-medoides

Il est introduit par Kaufman et Rousseeuw. L'esquisse de l'algorithme k-Medoides ressemble à celle de k-Means sauf que, contrairement à l'algorithme k-Means où la classe est représentée par une valeur moyenne, le centroïde, dans l'algorithme k-Medoids, une classe est représentée par un de ses objets prédominants, le Médoïde.

L'algorithme k-Medoids utilise une fonction objective qui définit la distance moyenne entre un objet et le médoïde.

Comme avantages de cet algorithme par rapport à k-Means, on cite :

- Il s'adapte à n'importe quel type de données ;
- Il est insensible aux objets isolés.

Pour améliorer la qualité de clustering de L'algorithme k-Medoids, de nombreuses variantes s'y sont succédées, entre autres les algorithmes PAM, CLARA et CLARANS.

Ces algorithmes adoptent le même principe de représentation par médoïde et utilisent-la même fonction objectif donnée ci-dessus.

2.6.3 Une méthode de clustering flou

Fuzzy C-Means (FCM) C'est une méthode ancienne

qui produit pour chaque document une probabilité d'appartenance à chaque topic. La somme de ces probabilités est égale à 1, alors que pour pLSA et LDA c'est la somme pour chaque topic, et pour NMF c'est la somme des carrés des coefficients de typicité pour chaque topic également qui doivent être égales à un. Les KMA pour leur part produisent pour chaque document des indicateurs de centralité dans le topic, à valeur entre 0 et 1 - ceci aussi pour les documents hors du topic d'appartenance. Pour les FCM il faut préciser au départ le nombre de topics désirés. Elles optimisent un critère basé sur des sommes de ces valeurs de probabilité à la puissance α , avec α supérieur à 1. Nous examinerons empiriquement plus bas si l'oubli dans lequel sont tombés les FCM était mérité ou pas... [11]

a. méthode Fuzzy C-means :

Fuzzy C-means (FCM) est une méthode de clustering qui permet à un objet de données d'appartenir à deux ou plusieurs clusters. Cette méthode dérivée de l'algorithme c-means [Ball et Hall, 1967], identique à l'algorithme k-means décrit précédemment, elle a été développée par Dunn [Dunn, 1973] en 1973 et améliorée par Bezdek [Bezdek, 1981] en 1981, est fréquemment utilisée dans la reconnaissance des formes. Il est basé sur la minimisation de la fonction objective suivante :

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - c_j\|^2 \quad 1 \leq m < \infty$$

Où m est un nombre réel (> 1), u_{ij} est le degré d'appartenance de x_i dans le j ème Cluster, x_i est le i ème élément des données mesurées, c_j est le centre d'un cluster et $k \cdot k$ est toute norme exprimant la similarité entre les données mesurées et le centre. Ce Partitionnement logique floue (fuzzy) est réalisé grâce à une optimisation itérative de la fonction objectif indiqué ci-dessus, avec la mise à jour de l'appartenance u_{ij} et les centres des clusters c_j .

On peut résumer la différence entre fuzzy C-means et k-means dans la fonction d'appartenance d'un nuage de points dans deux clusters dans l'exemple suivant :

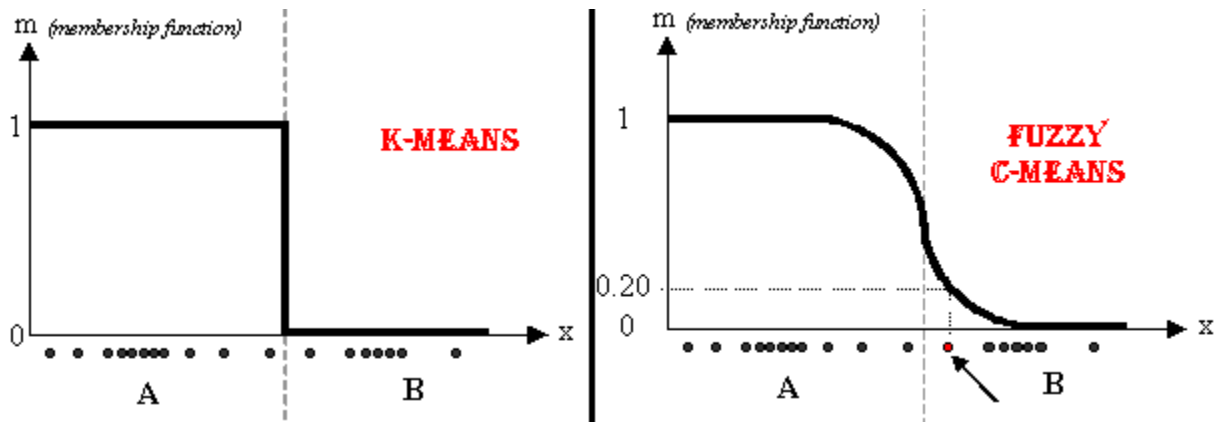


Figure 2.4 – Fonction d'appartenance dans kmeans/Fuzzy C-means

Dans le cas de k-means un objet ne peut pas appartenir dans deux clusters Simultanément, ce qui explique la Discrimination binaire entre les clusters mais en FCM il est possible qu'un objet appartienne à deux ou plusieurs clusters selon différents pourcentages que les données sont liés à chaque groupe par le biais d'une fonction d'appartenance, ce qui représente le comportement floue de cet algorithme. Pour le faire, nous devons simplement construire une matrice appropriée nommée U dont les facteurs sont des nombres entre 0 et 1, et représentent le degré d'appartenance entre les centres de données et des clusters.

$$U_{FCM} = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \\ 0.6 & 0.4 \\ \dots & \dots \\ 0.9 & 0.1 \end{bmatrix}$$

Il est également important de noter que les initialisations différentes causent différentes évolutions de l'algorithme. En fait, il pourrait converger vers le même résultat, mais probablement avec un nombre différent d'itérations

b. Algorithme

(Il ya des parties des équations caché il faut les récriture)

1. Initialiser $U = [u_{ij}]$ matrice $U_{(0)}$.
2. A la k-étape : calculer les centres $C_{(k)} = [c_j]$ avec $U_{(k)}$

$$c_j = \frac{\sum_{i=1}^N u_{ij}^m \cdot x_i}{\sum_{i=1}^N u_{ij}^m}$$

3. Mise à jour de $U_{(k)}$, $U_{(k+1)}$

$$u_{ij}^{j=1} = \frac{\|x_i - c_k\|^{-2m}}{\sum_{k=1}^m \|x_i - c_k\|^{-2m}}$$

4. Si $\|U_{(k+1)} - U_{(k)}\| < \varepsilon$ ($0 < \varepsilon < 1$), alors STOP, sinon le retour à l'étape 2.

Une méthode que son caractère hybride (la notion de centre de gravité et la notion Floue) le rend simple, rapide. La FCM exige des paramètres d'entrées, et que la matrice de partition floue, doit être initialisée d'une manière appropriée. Ces paramètres sont choisis d'une façon arbitraire, ces paramètres ont une grande influence sur le résultat attendu. Ce qu'il nous oblige de faire une étude approprié sur les données en entrée et le regroupement que l'on souhaite obtenir. Ce type d'algorithme est fort utilisé en traitement d'images.

2.6.4 Les algorithmes basés sur la densité

Ces algorithmes considèrent les classes comme étant des régions denses dans l'espace d'objets. Un objet de l'espace est dense si le nombre de ses voisins dépasse un certain seuil fixé à priori. Ils essayent alors d'identifier les classes en se basant sur la densité des objets dans une région. Les objets sont alors groupés non pas sur la base d'une distance mais sur la base de la densité de voisinage excède une certaine limite.

Parmi les algorithmes les plus connus dans cette catégorie, nous citons DBSCAN et OPTICS.

➤ L'algorithme DBSCAN

DBSCAN (*Density Based Spatial Clustering of Application with Noise*) est introduit par Ester, Kriegel, Sander et Xu. La démarche générale de DBSCAN se résume comme suit :

1. Sélectionner aléatoirement un objet p de l'ensemble d'objets X .
2. Déterminer l'ensemble E des objets accessible depuis p dans X .

3. Si p est un noyau, alors E est une classe.
4. Sinon sélectionner un autre objet et aller en (2).
5. S'arrêter lorsque tous les objets ont été sélectionnés.

Comme avantages de DBSCAN, on cite :

- Il peut chercher les classes de forme arbitraire ;
- Il est insensible à l'ordre d'arrivée des objets ;
- Il est incrémental car un nouvel objet qui arrive peut affecter certains voisinages.

Parmi les inconvénients de cet algorithme, on cite :

- Il est difficile à préciser le rayon de voisinage ϵ ;
- La performance peut être diminuée pour les données de haute dimension.

2.6.5 Les algorithmes basés sur la grille

Le principe de cette famille d'algorithmes est qu'on divise l'espace de données en cellules (fusion des régions jugées denses, selon la valeur d'un seuil fixé a priori, les régions jugées peu denses permettent d'établir des frontières.). Donc ce type d'algorithme est conçu pour des données spatiales. Une cellule peut être un cube, une région ou un hyper rectangle. En fait, elle est un produit cartésien de sous intervalles d'attributs de données.

Avec une telle représentation des données, au lieu de faire la classification dans l'espace de données, on la fait dans l'espace spatial en utilisant des informations statistiques des objets dans la cellule.

Les algorithmes les plus connus dans cette catégorie sont STING, CLIQUE et WaveCluster.

2.7 Techniques d'évaluations du clustering

L'évaluation des résultats d'un clustering est un problème majeur, qui renvoie à la question première : qu'est-ce qu'un bon schéma de clustering ? Cette problématique est plutôt bien synthétisée dans les récents travaux de M. Halkidi (77, 76). Trois cheminements sont alors envisagés pour évaluer ou comparer entre eux des schémas de clustering :

- 1- **L'évaluation externe** : il s'agit de confronter un schéma avec une classification prédéfinie. L'évaluation porte donc sur l'adéquation entre le schéma obtenu et une connaissance "externe" sur les données (schéma attendu).
- 2- **L'évaluation interne** : ce type d'évaluation n'utilise pas de connaissances externes mais uniquement les données d'entrées (matrice de (dis) similarité, descriptions des données etc.) comme référence. Ainsi, par exemple, parmi plusieurs schémas, le meilleur sera celui qui conserve un maximum d'information relativement à l'information contenue dans la matrice de (dis)similarité.

- 3- **L'évaluation relative** : cette dernière stratégie d'évaluation porte généralement sur les deux principaux critères de dispersions intra-clusters (à minimiser) et inter-clusters (à maximiser). L'évaluation relative est souvent utilisée pour comparer plusieurs schémas obtenus par une même méthode avec différents paramétrages. Ceci permet alors de sélectionner les paramètres optimaux pour un algorithme, étant donné un ensemble de données.

Il existe deux critères qui ont été largement considérés suffisants pour mesurer la qualité du partitionnement de données (Halkidi et al)

Compacité : Les objets situés dans un cluster doivent être similaires entre eux et différents des objets appartenant aux autres clusters. La variance des objets dans un cluster est un indice de compacité.

Séparation : les clusters doivent être bien séparés entre eux. La distance Euclidienne entre les centroïdes des clusters donne une indication sur le degré de séparation.

Dans cette section, nous décrivons les deux indices couramment utilisés pour évaluer la qualité du clustering la largeur de la silhouette et l'indice de Dunn. Ces mesures internes peuvent également être utilisées pour déterminer le nombre optimal clusters dans les données.

➤ **Coefficient de silhouette :**

L'analyse des silhouettes mesure comment une observation est regroupée et estime la distance moyenne entre les clusters. Le plot de la silhouette affiche une mesure de la distance entre chaque point dans un cluster et des points dans les clusters voisins.

Pour chaque observation i , la largeur de la silhouette s_i est calculée comme suit :

1. Pour chaque observation i , calcule la dissimilarité moyenne a_i entre i et tous les autres points du cluster auquel i elle appartient.
2. Pour tous les autres clusters C , à laquelle i n'appartient pas, calcule la dissimilarité moyenne $d(i, C)$ de i à toutes les observations de C . Le plus petit de ceux-ci $d(i, C)$ est défini par $b_i = \min_C d(i, C)$. La valeur de b_i peut être vue comme la dissimilarité entre i et son cluster «voisin», c'est-à-dire le cluster le plus proche auquel il n'appartient pas.
3. Enfin, la largeur de la silhouette de l'observation i est définie par la formule:

$$S_i = (b_i - a_i) / \max(a_i, b_i).$$

La largeur de la silhouette peut être interprétée comme suit :

- Les observations avec un grand S_i (presque 1) sont très bien regroupées.
- Un petit S_i (autour de 0) signifie que l'observation se situe entre deux clusters.
- Les observations avec un S_i négatif sont probablement placées dans le mauvais cluster.[12]

➤ Dunn index

L'indice Dunn est une autre mesure de validation de clustering interne qui peut être calculée comme suit:

Pour chaque cluster, calculer la distance entre chacun des objets du cluster et les objets des autres clusters

Utilisez le minimum de cette distance par paire comme séparation inter-cluster (séparation min.)

1. Pour chaque cluster, calculez la distance entre les objets du même cluster.
2. Utiliser la distance maximale (c.-à-d. le diamètre maximum) comme compacité intra-cluster
3. Utilisez la distance maximale intra-cluster (c'est-à-dire le diamètre maximal) comme compacité intra-cluster
4. Calculez l'indice de Dunn (D) comme suit:

$$D = \frac{\text{min. separation}}{\text{max. diameter}}$$

2.8 Caractéristiques des méthodes de clustering

- Mise l'échelle.
- Capacité à gérer différents types d'attributs.
- Découverte de clusters avec des formes arbitraires.
- Besoin minimum de connaissances du domaine pour déterminer les paramètres.
- Capacité à gérer le bruit et les exceptions.
- Indifférent à l'ordre des données en entrée.
- Nombre de dimensions.
- Incorporation de contraintes par l'utilisateur.
- Interopérabilité et utilisabilité.

2.9 Domaines d'applications du clustering

Le clustering possède des domaines d'applications extrêmement varies, parmi lesquels :

- 1- **Le Marketing** : segmentation du marché en découvrant des groupes de clients distincts à partir de bases de données d'achats.
- 2- **L'environnement** : identification des zones terrestres similaires dans une base de données contenant des informations (en termes d'utilisation) de la terre.

3- **les assurances** : identification de groupes d'assurés distincts associés à un nombre important de déclarations.

4- **La planification des villes** : identification de groupes d'habitations suivant leur type, valeur, localisation géographique.

5- **La médecine** : Localisation de tumeurs dans le corps humain.

Par exemple, dans un nuage de points fournis par le scan du cerveau, on identifie les points définissant une tumeur.

6- **La segmentation d'images** : Détection des zones homogènes dans une image.

7- **Web log analyses** : Identification de profils d'utilisateurs à travers leur flux de clics (Clickstream).

8- **Text mining** : Classification des textes selon leur similitude dans des dossiers automatiques.

2.10 Conclusion

Dans ce chapitre, Nous avons expliqué les méthodes de « clustering » qui est une des techniques statistiques largement utilisées dans la fouille de données. Il est dans un cadre d'apprentissage non supervisé, qui tente d'obtenir des informations sans aucune connaissance préalable, ce qui n'est pas le cas de l'apprentissage supervisé. Et nous présentons différents algorithmes de clustering et la problématique générale du clustering. Ces méthodes de clustering comme toutes les autres méthodes de classification, ont leurs avantages, faiblesses et définissent les trois étapes majeures de ce processus, à savoir : la préparation des données, le clustering proprement dit et l'exploitation des résultats de l'algorithme. Pour chacune de ces étapes, nous avons focalisé notre attention sur les grandes difficultés rencontrées, donnant lieu à des sous-domaines de recherche, comme par exemple : le choix d'une mesure de (dis)similarité ou d'une méthode de regroupement adaptée ou encore détermine le nombre optimal de clusters en clustering et l'évaluation et la validation des schémas de clustering obtenus.

Chapitre III :
Conception et
Implémentation

Chapitre 3 : Conception et Implémentation

3.1 Introduction

Lorsqu'on étudie simultanément un nombre important de variables quantitatives, comment en faire un graphique global ? La difficulté vient de ce que les individus étudiés ne sont plus représentés dans un plan, espace de dimension 2. C'est le fameux « curse of dimensionality ».

Les méthodes réduction des données sont essentielles pour analyser la quantité croissante de données de haute dimension. Les données de haute dimension sont souvent transformées en données de dimension inférieure via l'analyse en composantes principales (PCA), où des modèles cohérents peuvent être détectés plus clairement. Cette réduction de dimension non supervisée est utilisée dans différents domaines tels que l'analyse génomique et la récupération d'informations....etc. L'objectif de l'analyse en composantes principales est de revenir à un espace de dimension réduite en déformant le moins possible la réalité .

3.2 L'Analyse en Composantes Principales

a. Définition

L'analyse en composantes principales est une méthode factorielle de réduction de dimension pour l'exploration statistique de données quantitatives complexes dans un jeu de données. Exprimer ces informations sous la forme d'un ensemble de quelques nouvelles variables appelées composantes principales, construction du modèle statistique associé, estimation. Représentations graphiques des individus, des variables et simultanée, qualité de représentation. **Le but c'est on cherche à réduire le nombre de descripteurs (variables) avec le minimum de perte d'information et préservant les relations existant déjà avec entre les différents descripteurs .**

Mathématiquement, l'analyse en composantes principales est un simple *changement de base* : passer d'une représentation dans la base canonique des variables initiales à une représentation dans la base des facteurs définis par les vecteurs propres de la matrice des corrélations. L'objectif de l'analyse en composantes principales est de revenir à un espace de dimension réduite en déformant le moins possible la réalité. Il s'agit donc d'obtenir le résumé le plus pertinent des données initiales. [21]

b. Les principes

En PCA les données se présentent dans une matrice X à n lignes et p colonnes où

- chaque ligne représente un individu
- chaque colonne représente une variable
- Les variables sont quantitatives : la matrice X est constituée de valeurs numériques

Soit p variables statistiques réelles X^j ($j = 1, \dots, p$) observées sur n individus

i ($i = 1, \dots, n$) affectés des poids w_i :

$\forall i = 1, \dots, n : w_i > 0$ et

$\forall i = 1, \dots, n : x_i^j = X^j(i)$; mesure de X^j sur le $i^{\text{ème}}$ individu :

Ces mesures sont regroupées dans une matrice X d'ordre $(n \times p)$.

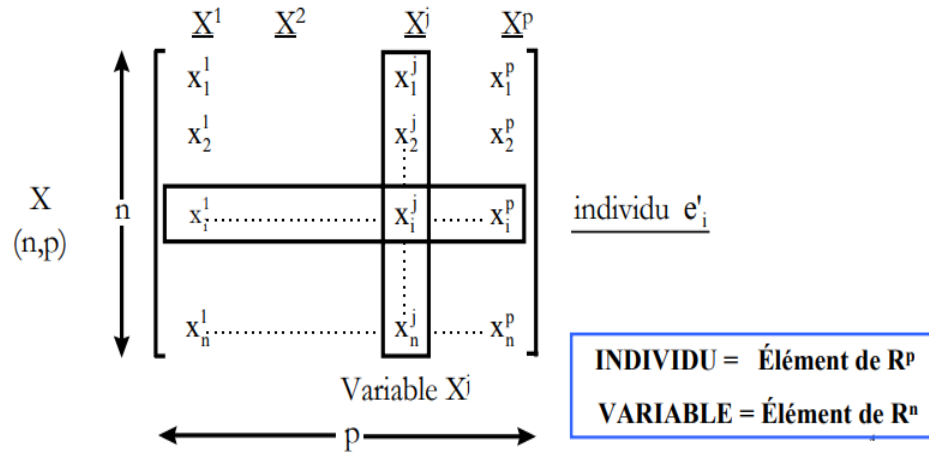


Figure 3.1 : Tableau de données en PCA

- À chaque individu i est associé le vecteur x_i contenant la i -ème ligne de X mise en colonne. C'est un élément d'un espace vectoriel noté E de dimension p ; nous choisissons $\mathbf{R}^p$ muni de la base canonique $\mathcal{E}$ et d'une métrique de matrice $\mathbf{M}$ lui conférant une structure d'espace euclidien : E est isomorphe à $(\mathbf{R}^p; \mathcal{E}; \mathbf{M})$; E est alors appelé espace des individus.

- À chaque variable X^j est associé le vecteur x^j contenant la j -ème colonne centrée (la moyenne de la colonne est retranchée à toute la colonne) de X . C'est un élément d'un espace vectoriel noté F de dimension n , nous choisissons $\mathbf{R}^n$ muni de la base canonique $\mathcal{F}$ et d'une métrique de matrice $\mathbf{D}$ diagonale des poids lui conférant une structure d'espace euclidien : F est isomorphe à $(\mathbf{R}^n, \mathcal{F}, \mathbf{D}) = \text{diag}(w_1, \dots, w_n)$: F est alors appelé espace des variables.

c. La réduction de la dimension

On cherche à réduire le nombre de descripteurs (variables) avec le minimum de perte d'information et préservant les relations existant déjà avec entre les différents descripteurs.

Les variables de départ sont remplacées par « **des vecteurs propres** » appelés Composantes Principales de la matrice de :

variance-covariance : mesure la liaison entre les différents descripteurs

$$\Sigma = (\text{cov}(X_i, X_j))_{i,j}$$

Ou de la matrice de **corrélation** : même chose que Σ sauf qu'il s'agit d'un paramètre sans dimension :

$$R = (R_{ij})_{i,j}$$

Généralement on s'arrête quand au moins 75% de la variance est expliquée par la variance cumulée par Composantes Principales (CP).

d. Un vecteur propre

i est une valeur propre de la matrice X si et seulement si :

$$X\mathbf{v} = i\mathbf{v}$$

- Le vecteur $\mathbf{v}$ dans la relation ci-dessus est appelé **vecteur associé à i**

Le nombre de valeurs propres, $i_1 > \dots > i_p$, est égal au nombre de lignes = nombre de colonnes de la matrice X

Important : La somme des valeurs propres de X est égale à la **variance** contenue dans l'ensemble des données

e. Expression des composantes principales

D'un point de vue pratique les composantes principales s'écrivent

$$F_j = I_1 X_1 + \dots + I_p X_p$$

C'est-à-dire que F_j est une combinaison linéaire des variables initiales $X_1, \dots, X_p$.

En plus de cet aspect calculatoire on doit pouvoir faire des affirmations sur la qualité de la réduction et la qualité de la représentation graphique.

f. Représentation graphique

Lorsque les différentes CP ont été trouvées on peut représenter les différentes variables et les différents individus dans le plan CP1, CP2 comme illustré ci-dessous

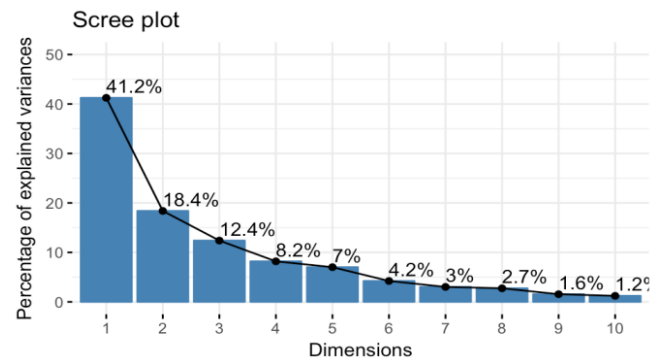


Figure 3.2 :Breast cancer : éboulis des dix premières valeurs propres de l'PCA

g. Interprétation

Chaque valeur propre représente la variance prise en compte par la composante principale correspondante. Pour l'exemple on obtient :

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6
standard deviation	3.6444	2.3857	1.67867	1.40735	1.28403	1.09880
Proportion of Variance	0.4427	0.1897	0.09393	0.06602	0.05496	0.04025
Cumulative Proportion	0.4427	0.6324	0.72636	0.79239	0.84734	0.88759
	PC7	PC8	PC9	PC10	PC11	PC12
standard deviation	0.82172	0.69037	0.6457	0.59219	0.5421	0.51104
Proportion of Variance	0.02251	0.01589	0.0139	0.01169	0.0098	0.00871
Cumulative Proportion	0.91010	0.92598	0.9399	0.95157	0.9614	0.97007

Figure 3.3 Interprétation de composant principale

Ici les deux premières composantes rendent compte de $0,5003 + 0,4917 = 0,9920 = 99,2\%$ de la variance totale.

Ce qui veut dire que les 4 descripteurs peuvent être remplacés par les 2 premières composantes tout en préservant la quasi-totalité de l'information (réduction).

h. Algorithme de PCA :**Algorithm :** PCA algorithm

- 1: **Input:** a D -dimensional training set $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ and the new (lower) dimensionality d (with $d \leq D$)
- 2: Compute the mean $\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$
- 3: Compute the covariance matrix $\text{Cov}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$
- 4: Find the spectral decomposition of $\text{Cov}(\mathbf{x})$, obtaining the eigenvectors $\xi_1, \xi_2, \dots, \xi_D$ and their corresponding eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_D$. Note that the eigenvalues are sorted, such that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D \geq 0$
- 5: For any $\mathbf{x} \in \mathbb{R}^D$, its new lower dimensional representation is:

$$\mathbf{y} = \left(\xi_1^T(\mathbf{x} - \bar{\mathbf{x}}), \xi_2^T(\mathbf{x} - \bar{\mathbf{x}}), \dots, \xi_d^T(\mathbf{x} - \bar{\mathbf{x}}) \right)^T \in \mathbb{R}^d,$$

and the original $\mathbf{x}$ can be approximated as

$$\mathbf{x} \approx \bar{\mathbf{x}} + (\xi_1^T(\mathbf{x} - \bar{\mathbf{x}}))\xi_1 + (\xi_2^T(\mathbf{x} - \bar{\mathbf{x}}))\xi_2 + \dots + (\xi_d^T(\mathbf{x} - \bar{\mathbf{x}}))\xi_d$$

3-3 R Studio

R est un langage de traitement et d'analyse de données de plus en plus répandu, notamment grâce à sa puissance et au fait qu'il est libre, gratuit et multiplateforme. est une suite logicielle permettant de manipuler des données sous forme de texte, de chiffres qu'ils soient géo référencés ou non, grâce aux nombreuses possibilités d'import depuis des sources très différentes, de les traiter, de les représenter sous forme graphique ou cartographique, puis de publier un rapport dynamique. Un IDE n'est pas une interface graphique au sens de SPSS ou Modalisa, qui permettrait d'utiliser le logiciel à travers des menus et des boîtes de dialogue : il s'agit d'un environnement facilitant la saisie, l'exécution de code, la visualisation des résultats, etc.

R Studio est multiplateforme, vous pouvez donc le télécharger et le faire fonctionner aussi bien sous Windows, Mac OS X ou Linux

Son interface se présente sous la forme d'une unique fenêtre découpée en quatre zones que l'on peut redimensionner, masquer ou maximiser selon ses préférences.[22]

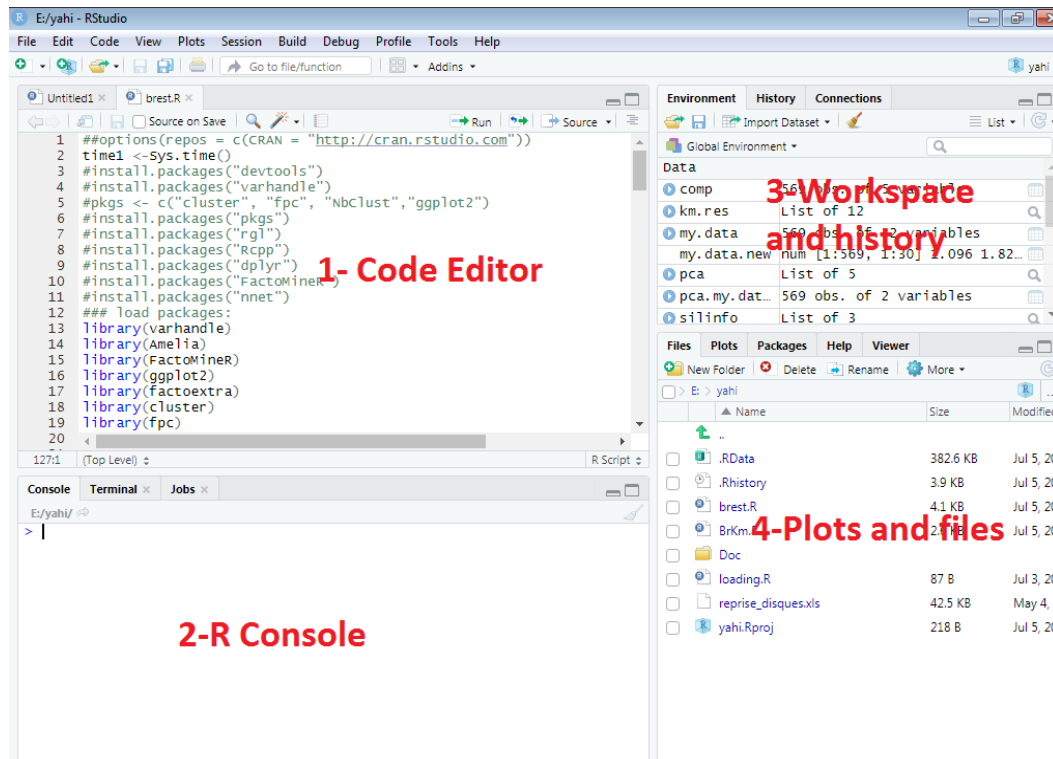


Figure 3.4 – L’interface de R Studio au démarrage

Une zone permet l’éditior de fichiers source R (avec coloration syntaxique et autocomplétion des noms de fonction et d’objet avec la touche *tabulation*), et une autre affiche la console avec la session R en cours d’exécution. L’utilisation du raccourci *Ctrl+Entrée* permet d’exécuter directement une ligne ou une sélection depuis le fichier source.

Une troisième zone permet de basculer entre l’affichage des objets de l’espace de travail actuel (*Workspace*) et l’historique des commandes exécutées. On peut même inspecter visuellement le contenu de certains objets :

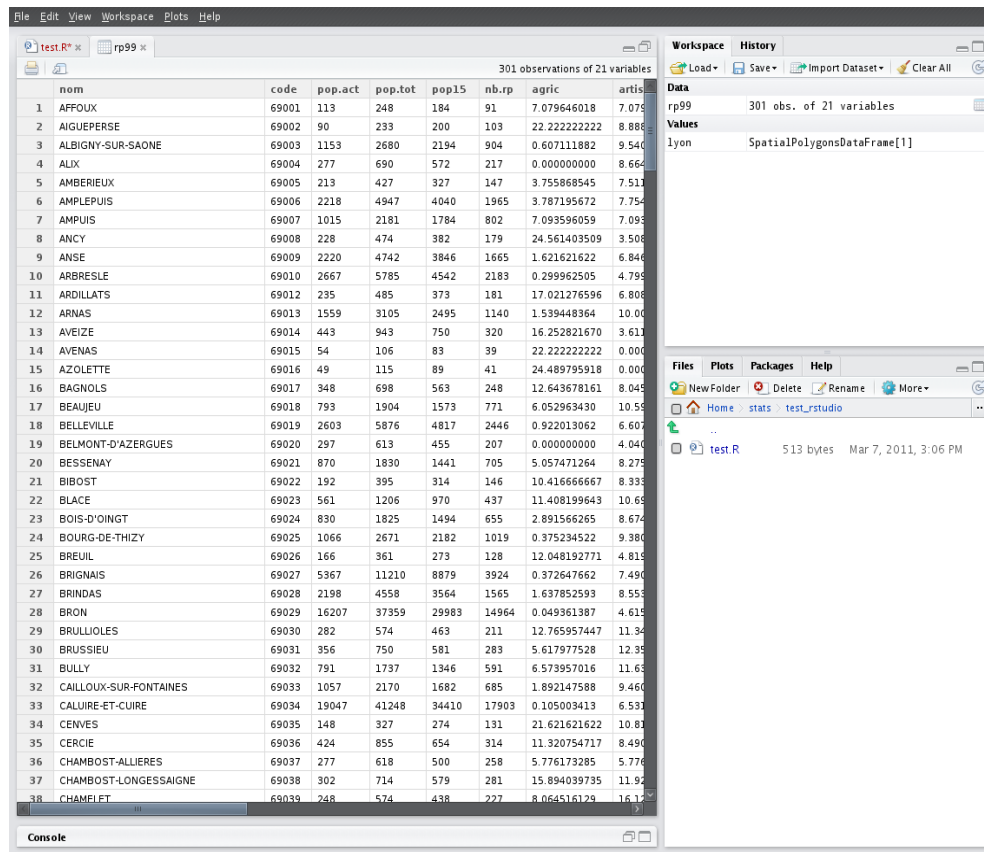


Figure 3.5 – troisième zone de R Studio

Enfin, une quatrième zone permet de basculer entre :

- un navigateur de fichiers
- la fenêtre d'affichage et d'export des graphiques
- une liste des extensions installées, qui permet de les charger en mémoire ou d'en installer de nouvelles très simplement
- un navigateur d'aide qui permet à la fois la navigation dans l'aide en ligne intégrée à R et l'affichage des pages d'aide des différentes fonctions

Au final on obtient une interface à la fois claire et complète, qui n'a pour seul inconvénient que d'être uniquement disponible pour le moment en anglais

`sessionInfo()` # Commande in R permet d'afficher les infos sur l'environnement

R version 3.4.4 (2018-03-15)

Platform: x86_64-w64-mingw32/x64 (64-bit)

Running under: Windows 7 x64 (build 7601) Service Pack 1....

3-4 Les résultats détaillés de clustering jeu de données Brest Cancer

a. Cancer du sein (32 variable / 569 observation) :

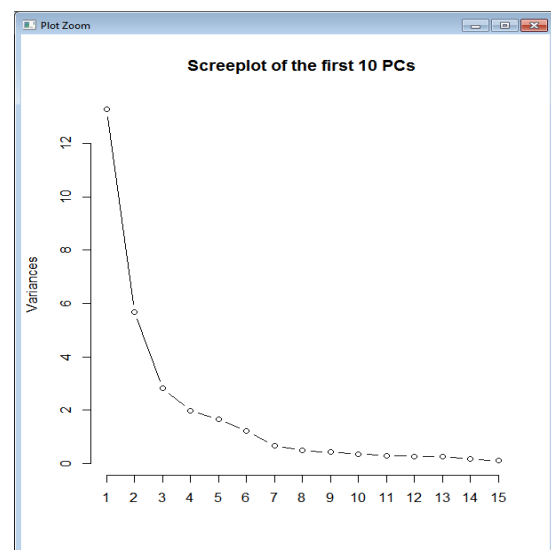
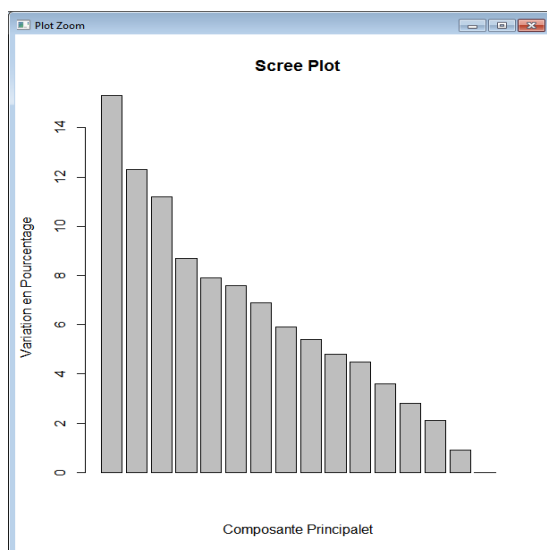
Jeu de données sur le cancer du sein comprend 569 observations et 33 variables. Les échantillons ont été divisés en deux catégories de diagnostic :

(**M** = malin, **B** = bénin)

J'ai d'abord fait une PCA et ai utilisé PC1 et PC2, car elles représentaient des valeurs propres cumulées de 63,3% (Dim1 44,3% + Dim2 19%) de la variance dans l'ensemble du jeu de données en utilisant simplement ces deux composantes principales, ce qui est plutôt bien compte tenu du fait que les données d'origine consistaient en 30 entités qu'il serait impossible de tracer de manière significative.

Ensuite, j'ai effectué K-means où j'ai utilisé $k = 2$ car il était capable de regrouper les données plus précisément.

Je vous présente ici les types de résultats que l'on a pu obtenir au cours de ces tests :



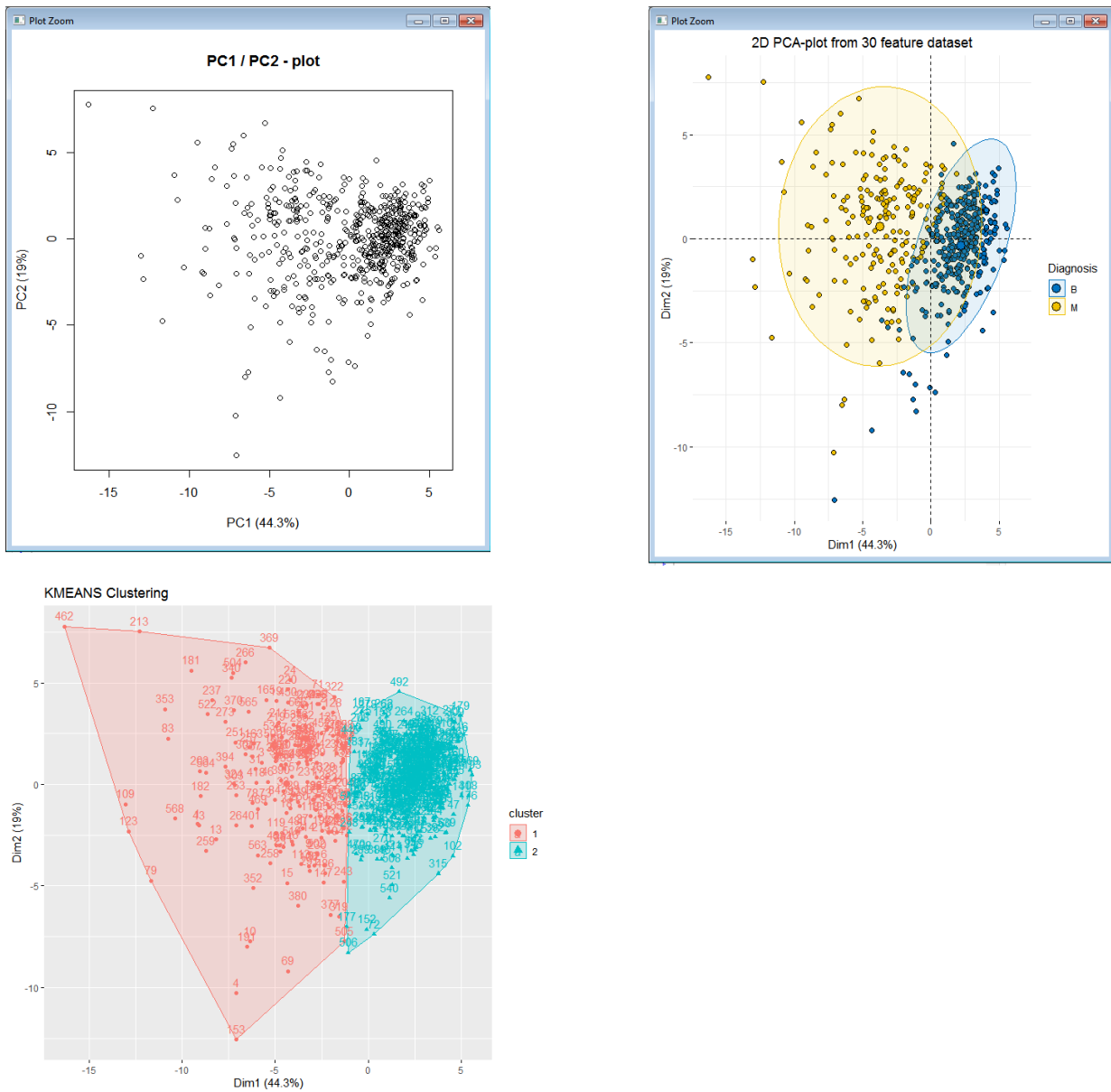


Figure 3.6 Different plot en R cluster PCAK-means (Data Set Breast Cancer)

3-5 Description d'Application

L'objectif de notre application est d'analyser et regrouper différents types de jeu de données microarray par la méthode de classification non-supervisée parmi la multitude des méthodes pouvant être utilisées, nous avons opté pour la méthode de clustering K-means et PCAk-means implémenté dans l'environnement R Studio .

3-6 La conception globale

Dans cette partie définit l'architecture globale de notre application sur R Studio .

pratiquement , il existe un grand nombre de clustering possibles Ici, nous n'allons en évoquer qu'une la méthode des « k-mean » « PCAk-mean » : Figure 3.3

3.6.1 Importer et préparation des données (Microarray)

Dans tout projet de clustering , une étape indispensable s'impose, la préparation des données. Une fois le « dataset » constitué, une phase de « data pre-processing » s'impose donc systématiquement. Cette préparation des données est indispensable pour tirer le meilleur des algorithmes de machine learning.

3.6.2 Normalisation des données

L'exploration de données peut générer des résultats efficaces si la normalisation est appliquée à jeu de données. C'est un processus utilisé pour normaliser tous les attributs du données et leur donner un poids équivalent afin d'éliminer les éléments redondants ou bruyants et de disposer de données valides et fiables qui améliorent la précision des résultats d'algorithme K-Means utilise la distance euclidienne très sujette aux irrégularités dans la taille de divers éléments Il existe différentes méthodes de normalisation des données telles que Min-Max, Z-Score et Décimal. La meilleure méthode de normalisation dépend des données à normaliser.

Normalisez les valeurs scalaires réelles des jeux de données avec les valeurs maximales et minimales à l'aide de l'équation:

$$v' = \frac{v - \min(e)}{\max(e) - \min(e)}$$

min (e) et max (e) sont les valeurs minimale et maximale de l'attribut E.

3.6.3 L'analyse en composantes principales (PCA)

Avant de vous lancer dans les k-means, gardez bien en tête leur forte sensibilité au grand nombre de variables, c'est le fameux « *curse of dimensionality* ». Il est courant d'appliquer Pca (analyse en composantes principales) avant un algorithme de classification (tel que k-means). pour réduction la dimension des données quantitatives complexes dans un jeu de données $X = (N \times M)$ à autre matrice $X' = (N' \times M')$ on pense que cela améliore les résultats de regroupement dans la pratique (réduction du bruit).(voir la page 40).

3.6.4 K-means

D'abord, vous devez paramétrer un nombre de centres de cluster 2, qui sont placés de manière aléatoire. Ensuite, chaque point de votre jeu de données est assigné au centre dont il est le plus proche. Chaque cluster reçoit un nouveau centroïde, selon une mesure de centralité, et chaque point est assigné au centroïde dont il est le plus proche. Ce processus est répété jusqu'à ce que la compacité et la séparation soit maximales (voir la page 27).

3.6.5 Validation de cluster

Il existe plusieurs méthodes pour comprendre la qualité d'un cluster. on a choisie les mesures externes pour calcule le taux de réussite

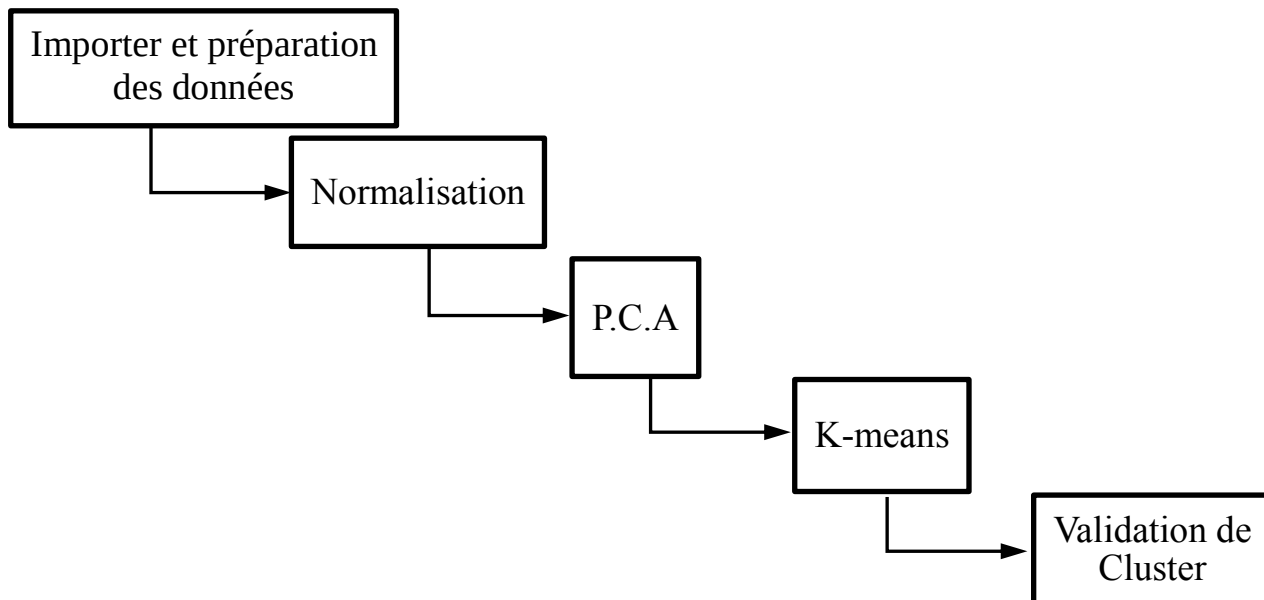


Figure 3.7: Schéma représentant la conception de l'approche proposé.

3.7 Descriptif des jeux de données réels pour le clustering

Ici, Quatre ensembles des jeu données différents sont considérés pour comparer les performances de la mise en cluster des k-means et PCA k-means.

Trois jeu de données proviennent du site [23]

Et un jeu de données (**Breast Cancer Wisconsin**) proviennent de site [24]

Le nombre d'échantillons varie de entre 32 à 70 et le nombre de gènes varie entre 569 à 7131 dans les données de microarray.

A.1 Breast Cancer Wisconsin

Jeu de données sur le cancer du sein comprend des mesures de $X(32 \times 569)$, deux Class

(M = malin, B = bénin)

A.2 leukemia

Jeu de données sur la leucémie du lymphome contient des mesures de l'expression génique de (71×7130) , deux Class (AML-ALL)

A.3 Colon

Jeu de données sur la cancer de colo contient des mesures de $X(62 \times 7130)$.

A.4 Central Nervous System (CNS)

Jeu de données sur les tumeurs du système nerveux central, y compris l'hypophyse contient des mesures de $X(60 \times 7131)$.

3.8 Résultats et discussions

À partir des résultats et des analyses effectuées et en fonction des objectifs de cette recherche, il peut être a conclu qu'entre k-means et PCAk-means, la comparaison la valeur de la meilleure validité de cluster c'est K-Means, toutes les expériences effectuées sont en appliquant deux clusters et deux composant principale (PC1 et PC2), k-means a l'avantage sur chaque expérience. Dans le cas de l'utilisation différent de quatre jeu de donnée, le résultat de la valeur du taux de réussite est bon pour k-means avec un bon temps d'exécution est donc K-Means est un algorithme optimal.

Cependant, PCA affecter la performance de cluster dans K-Means. Par conséquent, PCA K-Means il est clair que lorsque les dimensions sont réduites, les résultats s'améliorent systématiquement et de manière significative.

Les performances de quatre jeux de données sont comparées par k-mean vs PCAk-mean est représenté graphiquement sur la Figure 3.7

Les scores de l'indice ont été rapportés dans les tableau 3.2 respectivement :

Data Set	K-means		K-means + PCA	
	Taux de réussite %	Temps d'exécution	Taux de réussite %	Temps d'exécution
Breast Cancer	91.03	9.73 sec	90.68	20.02sec
Colon	54	17.22 sec	54	28.85sec
CNS	55	46.52 sec	55	1 min
Eukemila	75	44.17sec	73.61	56.94sec

Table 3.2. Comparaison taux de réussite K-means & PCA K-means

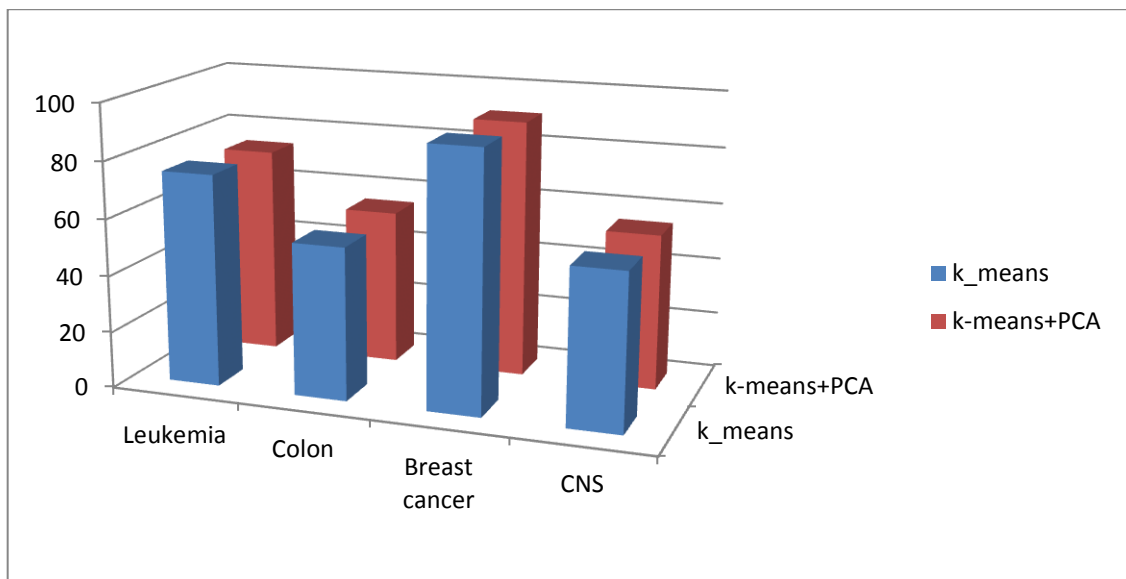


Figure 3.8 Performances de quatre jeux de données K-mean & PCAk-mean

3.9 Conclusion

Après simulation sur les différents exemples de data Set, on a confirmé que l'algorithme de clustering k-means avec l'analyse en composantes principales (PCA) est efficace pour grouper les données de la matrice microarray multidimensionnelle qui donne un bon taux de réussite et où tous les membres de chaque cluster sont proches les uns des autres. Les données de haute dimension sont souvent transformées en données de dimension inférieure via l'analyse en composantes principales (PCA) et pour visualiser les modèles dans le jeu de données, nous avons réussi ce que l'on appelle le fameux « curse of dimensionality ».

Conclusion Générale

Conclusion générale

Dans ce mémoire, nous avons abordé la problématique de la classification automatique des données microarray avec une méthode de classification non-supervisée, parmi la multitude des méthodes non-supervisée pouvant être utilisées, nous avons opté pour la méthode de clustering K-means et k-means avec L'analyse en composantes principales (PCA) une méthode très populaire et très performante par rapport aux autres.

Notre méthode de clustering à été testé sur des jeux de données diverses, Nous avons constaté à travers nos différents tests que les résultats obtenus montrent que la méthode de clustering PCAK-means apporte des résultats satisfaisantes notamment la réduction de dimensions de notre data Set .Nous avons implémenté une application sur l'environnement R Studio qui permet de regrouper des données de microarray. nous aient permis d'obtenir des résultats compétitifs, notre méthode pourrait être encore améliorées.

Bibliographie

- [01] <http://www.bioinfo4arabs.com>(consulté le 12/12/2018) (« notre traduction »)
- [02] Nancie Reymond- Bioinformatique des puces à ADN et application à l'analyse du transcriptome de *Buchnera aphidicola* -page 89.
- [03] <http://biochimej.univ-angers.fr/Page2/COURS/9ModulGenFoncVeg/5MethEtudGenFonc/3PucesADN/1PucesADN.htm> (consulté le 18/01/2019)
- [04] : <http://biotech-btk.wifeo.com/les-puces.php> (consulté le 24/01/2019) .
- [05] <https://onconano2.webnode.fr/diagnostique/puce-%C3%A0-adn/experience-1/>(consulté le 26/01/2019) .
- [06] J. HAN & M. KAMBER, Data mining: Concepts and techniques, Management Systems (The Morgan Kaufmann Series in Data Management Systems), MORGAN KAUFFMAN, ISBN 1-55860-901-6,
- [07] 2006Guénaél Cabanes, classification non supervisée à deux niveaux guidée par le voisinage et la densité, 03/12/10, Université Paris 13 pages 1
- [08] J. Hartigans. Clustering algorithms. John Wiley and Sons, Inc., 1975.
- [09] Guillaume Cleuzieu, Une méthode de classification non-supervisée pour l'apprentissage de règles et la recherche d'information, décembre 2004, Université d'Orléans page 07, pages11-15
- [10] Guillaume Cleuzieu une méthode de classification non-supervisé pour l'apprentissage de règle et la recherche d'information, université d'orléans, 2004
- [11] D. Bowtell and J. Sambrook DNA Microarrays. A Molecular Cloning Manual, ed 2003, New York. Cold Spring Harbor Laboratory Press. 712.
- [12] D.L. DAVIES & D.W. BOULDIN, A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 224-227, 1979
- [13] Y. BATISTAKIS & M.VAZIRGIANNIS M. HALKIDI, Clustering validity checking methods, SIGMOD Record, 2001.
- [14] <https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203345-clustering-definition/> (consulté le 02/03/2019)
- [15] Osmar R. Zaïane: “Principles of Knowledge Discovery in Databases - Chapter 8: Data Clustering”
- [16] <http://www.cs.ualberta.ca/~zaiane/courses/cmput690/slides/Chapter8/index.html>

(consulté le 08/04/2019)

[17] MEMOIRE Une approche multi agents pour Le Data Mining. Présenté par Kateb Nabila Université Larbi Ben M'hidi Oum El Bouaghi.

[18] J.C. DUNN, Well separated clusters and optimal fuzzy-partitions, Journal of Cybernetics, pages 95-104, 1974.

[19] The Chipping Forecast. Nat Genet, 2002. 32 Suppl: p. 461-552.

[20] <https://quanti.hypotheses.org/488/> (Consulte le : 05-07-2019)

[21] [https://www.news-medical.net/life-sciences/DNA-microarray-\(French\).aspx](https://www.news-medical.net/life-sciences/DNA-microarray-(French).aspx)(consulté le 05/05/2019)

[22] Site The MGuide. Version 2.0 : <http://cmgm.stanford.edu/pbrown/mguide/index.html> (consulté le 08/03/2019)

[23] <http://csse.szu.edu.cn/staff/zhuzx/Datasets.html>. (consulté le 26/06/2019)

[24]:<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>(consulté le 16/04/2019)

[25] <https://www.datanovia.com/en/lessons/cluster-validation-statistics-must-know-methods/#silhouette-coefficient>(consulté le 08/05/2019)

[26] <http://wikistat.fr/pdf/st-m-explo-intro.pdf> Consulte le : 02-07-2019

Liste des abréviations :

ADN	Acide Désoxyribonucléique
ADNc	ADN complémentaire
ARN	Acide Ribonucléique
ARNm	ARN massager
ARNr	ARN ribosomique
Cy5	Cyanine 5
Cy3	Cyanine 3
P32	Le Phosphore 32
PCR	Polymerase Chain Reaction
PCA	L'analyse en composantes principales

Liste des tableaux

1.1 Principaux types de puces à ADN.....	11
3.2 Comparaison taux de réussite K-means & K-means+ PCA	48

Liste des figures

Figure 1.1. Le principe de la fonction cellulaire (Central Dogma).....	03
Figure 1.2. Nucléotides et ADN sous sa forme double brin	04
Figure 1.3. Illustration de la structure de l'ADN et de l'ARN messenger	05
Figure 1.4. Une illustration du principe de coupe afin de créer différentes protéines à partir d'un gène....	06
Figure 1.5 : De l'ADN à l'ARN messenger (ARNm)	07
Figure 1.6 : Principe général de la traduction de l'ARNm en protéine.	07
Figure 1.7 : La première puce à ADN.....	09
Figure 1.8 : Les puces à oligonucléotides synthétisés in situ (« genechips »Affymetrix).....	12
Figure 1.9 : Principe des puces à ADN.....	14
Figure 2.1 : Les différentes étapes du clustering.....	18
Figure 2.2 : Exemple du clustering hiérarchique.....	24
Figure 2.3 : Exemple sur l'algorithme K-mean	26
Figure 3.1 : Tableau de données en PCA.....	38
Figure 3.2 : Breast cancer : éboulis des dix premières valeurs propres de l'PCA.....	39
Figure 3.3 Interprétation de composant principale	39
Figure 3.4 : L'interface de RStudio au démarrage.....	41
Figure 3.5 : troisième zone de R Studio	42
Figure 3.6 Different plot en R cluster K-means (Data Set Breast Cancer).....	44
Figure 3.7: Schéma représentant la conception de l'approche proposé.....	46
Figure 3.8 : Performances de quatre jeux de données K-mean & PCA k_mean	48

ملخص - تصف هذه المذكرة تحليل المجموعات لبيانات التعبير الجيني للميكرو أري و الحد من ابعاد مصفوفة للميكرو أري أهم مجال لتكنولوجيا ميكرو أري هو تحليل تجميع البيانات. يشير تحليل الكتلة إلى تقسيم مجموعة بيانات معينة إلى مجموعات بناءً على الميزات المحددة بحيث تكون نقاط البيانات داخل المجموعة أكثر تشابهاً مع بعضها البعض من النقاط في المجموعات المختلفة. في هذه المذكرة. K-mean مع تطبيق تحليل المكون الرئيسي (PCA) تعطي نتائج جيدة قريبة من نتائج خوارزمية K-mean.

الكلمات الرئيسية - التجميع, Kmeans, تحليل المكون الرئيسي, المصفوفة الدقيقة

Summary - this thesis describes the automatic classification analysis of microarray data, reduce the dimensions of the microarray. The most important area of microarray technology is data cluster analysis. Cluster analysis refers to the partition of a dataset to give a group based on specified features, so that the data points of a group are more similar to each other than the points of different groups. In this thesis, a classification algorithm by K-means with principal component analysis (PCA) that give good results equivalent to the result of k-means.

Keywords—Clustering, K-means, PCA, Microarray data.

Résumé- ce mémoire décrit l'analyse de classification automatique des données de puces à ADN, réduire les dimensions de la matrice puce à ADN. Le domaine le plus important de la technologie des puces à ADN est l'analyse par regroupement de données. L'analyse de cluster fait référence à la partition d'un jeu de données donner un groupes basés sur des fonctionnalités spécifiées, de sorte que les points de données d'un groupe soient plus similaires les uns aux les autres que les points de différents groupes. Dans ce mémoire, un algorithme de classification par K-moyennes avec l'analyse en composantes principales (ACP) qui a donner des bonnes résultat équivalent de la résultat de k- moyennes.

Mots-clés – Clustering, K- moyennes, PCA, puces à ADN