الجمهورية الجزائرية الديمقراطية الشعبية
RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
وزارة التعليم العالي و البحث العلمي
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
جامعة محمد بوضياف بالمسيلة
UNIVERSITE MOHAMED BOUDIAF DE M'SILA
كلية الرياضيات و الإعلام الآلي
Faculte Des Mathematiques Et De L'informatique

MEMOIRE DE MASTER

Option :

SYSTEME D'INFORMATION ET GENIE LOGICIEL

THEME

Techniques de Fouille de données appliquées Aux données massives (Big data) Étude comparative

Présenté par : -BOUCENNA TOUFIK
- YAHIAOUI OMAR

Encadré par :
Dr. LAMICHE CHAABANE

Co-Encadreur:
Dr. MEHENNI TAHAR

Promotion : 2019-2020

Remerciements

Tout d'abord nous remercions Allah Le Tout Puissant de nous avoir donné la

force, la volonté, et le privilège d'étudier et de suivre

le chemin de la science;

À l'âme du défunt, le **Dr :LAMICHE CHAABANE Rahimahou Allah**

Nous remercions les membres du jury, d'avoir accepté de porter un jugement

sur ce travail

Nous tenons également à remercier plus particulièrement :

Notre co-encadreur **Dr : MEHENNI TAHAR**

Tous les professeurs du département d'informatique, à qui l'on doit tout le respect.

Nos remerciements vont également à nos amis qui partagent avec nous

les bons moments de l'étude

Un grand merci, pour tous ceux qui ont, en quelque part, de près

ou de loin, participé à la réalisation

de ce travail.

Merci à tous

Dédicace

Je dédie ce travail à :

- *À mes très chers parents, que Allah Le Tout Puissant les protège*
- *À ma femme et sa famille*
- *À mes chers enfants*
- *À mes frères et mes sœurs*
- *À tous mes enseignants*
- *À mes amis ainsi que tous les gens qui m'ont aidé de près ou de loin à accomplir ce travail*
- *Tous mes collègues de travail.*

M : *Boucenna toufik*

Dédicace

Je dédie ce travail à :

- *Ma mère qui m'a indiqué que la volonté fait toujours les grands hommes et les grandes femmes.*
- *À mes frères et mes sœurs*
- *À tous mes enseignants*
- *À mes amis ainsi que tous les gens qui m'ont aidé de près ou de loin à accomplir ce travail*
- *Tous mes collègues de travail.*

M : YAHIOUI OMAR

Sommaire

Liste des figures

liste des tableaux

Introduction Générale

La fouille de données peut être définie comme le « Processus d'extraction non triviale d'informations implicites, inconnues auparavant et potentiellement utiles (sous forme de règles, contraintes, régularités) à partir de données issues de bases de données » (Gregory Piatetsky-Shapiro). Si ce domaine est loin d'être nouveau, c'est seulement depuis quelques années que les praticiens se confrontent à de nouvelles difficultés liées à une augmentation significative du volume de données. Cette augmentation a été dans certains cas bien plus rapide que la croissance continue des capacités de calcul et de stockage des serveurs individuels, les volumes résultants étant alors incompatibles avec un traitement centralisé. On parle alors en général de « données massives » (*big data*).

Problématique : Comment appliquer les techniques de fouille de données aux données massives (Big Data) .

L'objectif : Le but de notre étude appliquer quelques techniques clustering (k-means et hiérarchique) de fouille de données sur des données massives (et en faire une comparaison)

Ce mémoire s'articule autour de quatres chapitres comme suit:

Le Premier Chapitre consacré à la fouille de données, ses principes, ses méthodes et ses techniques

Le Deuxième Chapitre consacré aux données massives (big data)

Le Troisième Chapitre consacré aux technique de fouille de données sur des données massives à savoir le clustering (k-means et hiérarchique)

Le dernier chapitre présente l'expérimentation et la validation de système réalisé.

Enfin, nous concluons ce projet par une conclusion générale.

Chapitre I Fouille de Données

I-1 Introduction

Les données brutes, malgré leur quantité qui augmente d'une façon exponentielle, n'ont presque aucune valeur, ce qui est le plus important en fait c'est les connaissances pour lesquelles nous sommes tous assoiffés et qui sont obtenus par la compréhension de ces Données, mais plus on a des données plus ce processus devient difficile.

De nos jours, les changements de notre environnement sont dénotés par des capteurs qui sont devenus de plus en plus nombreux. Par conséquent, la compréhension de ces données est très importante. Et comme il est dit par Piatestky-Shapiro, *« […] as long as the world keeps producing data of all kinds […] at an ever increasing rate, the demand for data mining will continue to grow»* **[1]**. D'où la fouille de données devient une nécessité.

I-2 Définition Fouille De Données

Selon le Groupe Gartner, le Data Mining appelé aussi fouille de données est le processus de découverte de nouvelles corrélations, modèles et tendances en analysant une grande quantité de données, en utilisant les technologies de reconnaissance des formes ainsi que d'autres techniques statistiques et mathématiques**[2]**.

Ils existent d'autres définitions :

Le Data Mining est l'analyse de grandes ensembles de données observationnelles pour découvrir des nouvelles relations entre elles et de les reformuler afin de les rendre plus utilisables de la part de ses propriétaires **[3]**.

Le Data Mining est un domaine interdisciplinaire utilisant dans le même temps des techniques d'apprentissage automatiques, de reconnaissance des formes, des statistiques, des bases de données et de visualisation pour déterminer les manières d'extraction des informations de très grandes bases de données **[4]**. Le Data Mining est un processus inductif, itératif et interactif dont l'objectif est la découverte de modèles de données valides, nouveaux, utiles et compréhensibles dans de larges Bases de Données **[5].**

I-3 Intérêt Du Fouille De Données

Les entreprises sont inondées de données (scanners des supermarchés, internet, bases de données, etc.). Ces données languissantes dans des entrepôts de données (ou référentiels, ou *data warehouse*). **[6]**

Le data mining permet :

- ✓ d'exploiter ces données pour améliorer la rentabilité d'une activité.
- ✓ augmenter le retour sur investissement des systèmes d'information.

- ✓ produire de la connaissance et cela dans le but de comprendre les phénomènes dans un premier temps : SAVOIR et dans le but de prendre des décisions dans un second temps : PREVOIR pour DECIDER.

I-4 Processus Du Fouille De Données

Il est très important de comprendre que le data mining n'est pas seulement le problème de découverte de modèles dans un ensemble de donnée. Ce n'est qu'une seule étape dans tout un processus suivi par les scientifiques, les ingénieurs ou toute autre personne qui cherche à extraire les connaissances à partir des données. En 1996 un groupe d'analystes définit le data mining comme étant un processus composé de cinq étapes sous le standard CRISP-DM (Cross-Industry Standard Process for fouille de données) comme schématisé ci-dessous :

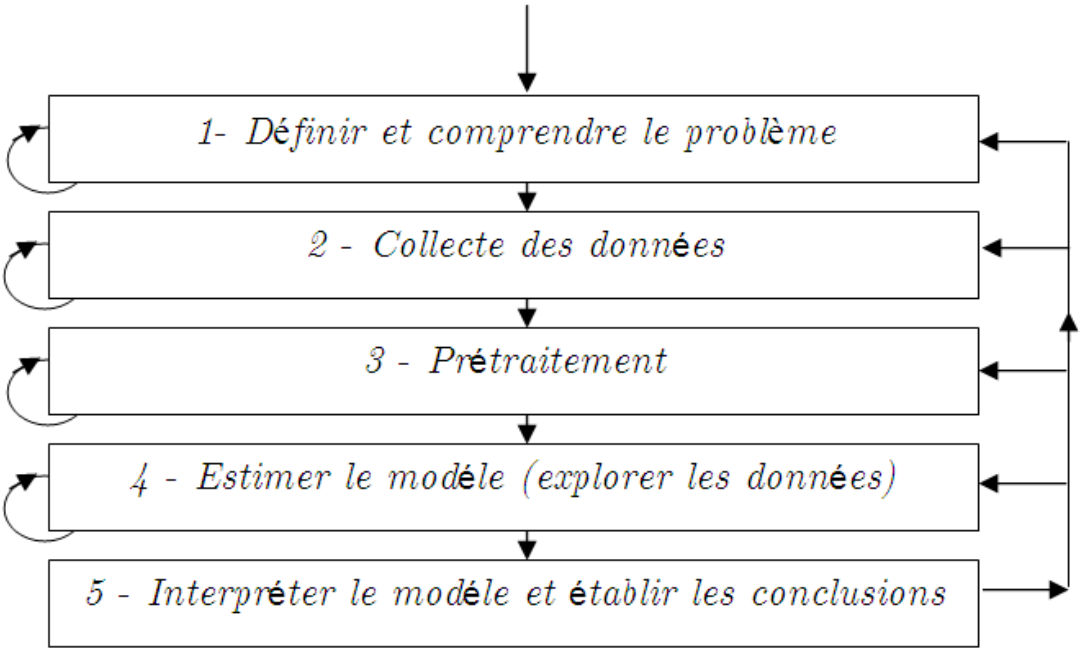

Figure. I-1 –Schéma d'un processus de fouille de données

Ce processus, composé de cinq étapes, n'est pas linéaire, on peut avoir besoin de revenir à des étapes précédentes pour corriger ou ajouter des données. Par exemple, on peut découvrir à l'étape d'exploration (5) de nouvelles données qui nécessitent d'être ajoutées aux données initiales à l'étape de collection (2).

I.4-1 Définition et Compréhension du Problème :

Dans la plus part des cas, il est indispensable de comprendre la signification des données et le domaine à explorer. Sans cette compréhension, aucun algorithme ne va donner un résultat fiable. En effet, Avec la compréhension du problème, on peut préparer les données nécessaires à l'exploration et interpréter correctement les résultats obtenus. Généralement, le data mining est

effectué dans un domaine particulier (banques, médecine, biologie, marketing, ...etc) où la connaissance et l'expérience dans ce domaine jouent un rôle très important dans la définition du problème, l'orientation de l'exploration et l'explication des résultats obtenus. Une bonne compréhension du problème comporte une mesure des résultats de l'exploration, et éventuellement une justification de son coût. C'est-à-dire, pouvoir évaluer les résultats obtenus et convaincre l'utilisateur de leur rentabilité.

I-4-2 Collecte Des Données :

Dans cette étape, on s'intéresse à la manière dont les données sont générées et collectées. D'après la définition du problème et des objectifs du data mining, on peut avoir une idée sur les données qui doivent être utilisées. Ces données n'ont pas toujours le même format et la même structure. On peut avoir des textes, des bases de données, des pages web, ...etc. Parfois, on est amené à prendre une copie d'un système d'information en cours d'exécution, puis ramasser les données de sources éventuellement hétérogènes (fichiers, bases de données relationnelles, temporelles, ...).

Quelques traitements ne nécessitent qu'une partie des données, on doit alors sélectionner les données adéquates. Généralement les données sont subdivisées en deux parties : une utilisée pour construire un modèle et l'autre pour le tester. On prend par exemple une partie importante (suffisante pour l'analyse) des données (80 %) à partir de laquelle on construit un modèle qui prédit les données futures. Pour valider ce modèle, on le teste sur la partie restante (20 %) dont on connaît le comportement.

I-4-3 Prétraitement :

Les données collectées doivent être "préparées" [?]. Avant tout, elles doivent être nettoyées puisqu'elles peuvent contenir plusieurs types d'anomalies : des données peuvent être omises à cause des erreurs de frappe ou à causes des erreurs dues au système lui-même, dans ce cas il faut remplacer ces données ou éliminer complètement leurs enregistrements. Des données peuvent être incohérentes c-à-d qui sortent des intervalles permis, on doit les écarter où les normaliser. Parfois on est obligé à faire des transformations sur les données pour unifier leur poids. Un exemple de ces transformations est la normalisation des données qui consiste à la projection des données dans un intervalle bien précis [0,1] ou [0,100] par exemple. Un autre exemple est le lissage des données qui considère les échantillons très proches comme étant le même échantillon.

Le prétraitement comporte aussi la réduction des données [?] qui permet de réduire le nombre d'attributs pour accélérer les calculs et représenter les données sous un format optimal pour l'exploration. Une méthode largement utilisée dans ce contexte, est l'analyse en composantes principales (ACP). Une autre méthode de réduction est celle de la sélection et

suppression des attributs dont l'importance dans la caractérisation des données est faible, en mesurant leurs variances. On peut même réduire le nombre de données utilisées par le data mining en écartant les moins importantes.

Dans la majorité des cas, le pré-traitement doit préparer des informations globales sur les données pour les étapes qui suivent tel que la tendance centrale des données (moyenne, médiane, mode), le maximum et le minimum, le rang, les quartiles, la variance, ... etc. Plusieurs techniques de visualisation des données telles que les courbes, les diagrammes, les graphes,... etc, peuvent aider à la sélection et le nettoyage des données. Une fois les données collectées, nettoyées et prétraitées on les appelle entrepôt de données (data warehouse).

I-4-4 Estimation Du Modèle :

Dans cette étape, on doit choisir la bonne technique pour extraire les connaissances (exploration) des données. Des techniques telles que les réseaux de neurones, les arbres de décision, les réseaux bayésiens, le clustering, ... sont utilisées. Généralement, l'implémentation se base sur plusieurs de ces techniques, puis on choisit le bon résultat. Dans le reste de ce rapport on va détailler les différentes techniques utilisées dans l'exploration des données et l'estimation du modèle.

I-4-5 Interprétation Du Modèle Et Etablissement Des Conclusions :

Généralement, l'objectif du data mining est d'aider à la prise de décision en fournissant des modèles compréhensibles aux utilisateurs. En effet, les utilisateurs ne demandent pas des pages et des pages de chiffres, mais des interprétations des modèles obtenus. Les expériences montrent que les modèles simples sont plus compréhensibles mais moins précis, alors que ceux complexes sont plus précis mais difficiles à interpréter.

I-5 Méthodes De Fouille De Données

Il existe plusieurs méthodes de fouille de données utilisées pour différents buts. On distingue deux types: les méthodes orientées vérification et les méthodes orientées découverte comme illustré dans la **figure I.2.**

I-5-1 Les Méthodes Orientées Découverte :

Ces méthodes identifient automatiquement les configurations à partir des données. Elles consistent en méthodes de prédictions et de descriptions.

Les méthodes descriptives sont orientées a` l'interprétation de données tel que le clustering, la summarisation, la visualisation etc. Les méthodes prédictives visent à établir Automatiquement un modèle comportemental qui obtient de nouveaux et invisibles échantillons et peut prévoir les valeurs d'une ou plusieurs variable Liées à l'´echantillon. Il

développe également les configurations, qui forment la connaissance découverte d'une manière compréhensible et facile à utiliser. Parmi ces méthodes on trouve celles qui se basent sur la régression et d'autres qui se basent sur la classification tel que les arbres de décision, les réseaux de neurones, etc.

La plus parts des méthodes orientées découverte (en particulier quantitatives) sont basées sur l'apprentissage inductif ou le modèle est construit explicitement ou implicitement par généralisation à partir d'un ensemble d'exemples.

I-5-2 Les Méthodes Orientées Vérification :

Elles procèdent en évaluant des hypothèses proposées par une source externe (expert, ...). Ces méthodes incluent les méthodes statistiques traditionnelles tel que le test d'hypothèses (t-test des moyennes), analyse de variance etc. Ces méthodes sont moins associées à la fouille de données par rapport aux méthodes orientées découverte, car la plus part des méthodes de fouille de données sont concernées par découvrir une hypothèse plutôt que tester celle qui sont déjà connues [1] .

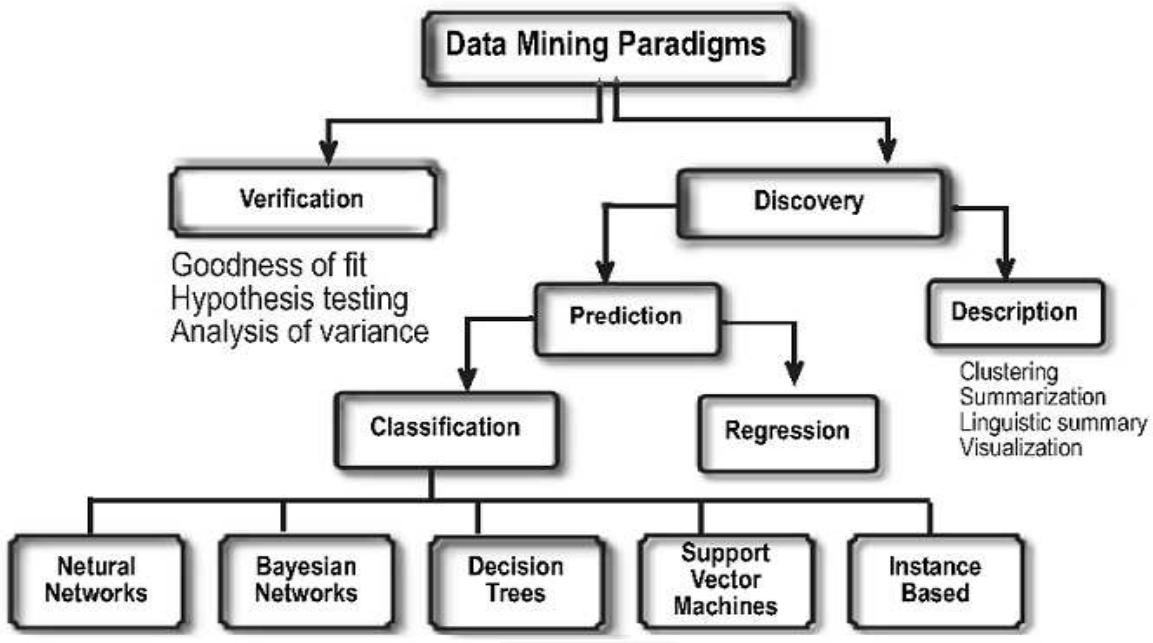

Figure. I-.2 Méthodes de fouilles de données

I-6 Composantes Des Algorithmes De Fouille De Données

Trois composantes principales peuvent être identifié´s dans un algorithme de fouille de données[6]:

- **Modèle De Représentation** : est le langage utilisé pour décrire les formes à découvrir. Si la représentation est trop limitée, alors le temps d'apprentissage est nul ou bien les exemples peuvent construire un modèle précis pour les données. Il est important qu'un analyste de données comprend entièrement les prétentions représentatives qui pourraient être inhérentes dans une méthode particulière. Il est également important qu'un concepteur d'algorithme formule clairement quelles prétentions représentatives sont faites par un algorithme particulier.
- **Modèle D'´Evaluation**: les critères d'´evaluation sont des rapports quantitatifs (ou une fonction de fitness) qui d´écrivent de combien un modèle particulier se rapproche du processus de découverte de connaissance.
- **Méthode De Recherche** : elle se compose de deux éléments les paramètres de recherche et le modèle de recherche.

I-7 Techniques Du Fouille De Fouille De Données

Pour effectuer les taches du Data Mining il existe plusieurs techniques issus de disciplines scientifiques diverses (statistiques, intelligence artificielle, base de données) afin de faire apparaître des corrélations cachées dans des gisements de données pour construire des modèles à partir de ces données. Dans ce chapitre, nous présentons les techniques du data mining les plus connues.

I-7-1 Les Réseaux De Neurones

Un réseau de neurones est un modèle de calcul dont le fonctionnement vise à simuler le fonctionnement des neurones biologiques, il est constitué d'un grand nombre d'unités (neurones) ayant chacune une petite mémoire locale et interconnectées par des canaux de communication qui transportent des données numériques. Ces unités peuvent uniquement agir sur leurs données locales et sur les entrées qu'elles reçoivent par leurs connections. Les réseaux de neurones sont capables de prédire de nouvelles observations (sur des variables spécifiques) à partir d'autres observations (soit les même ou d'autres variables) après avoir exécuté un processus d'apprentissage sur des données existantes.

La phase d'apprentissage d'un réseau de neurones est un processus itératif permettant de régler les poids du réseau pour optimiser la prédiction des échantillons de données sur Les quelles l'apprentissage été fait. Après la phase d'apprentissage le réseau de neurones devient capable de généraliser **[7]**.

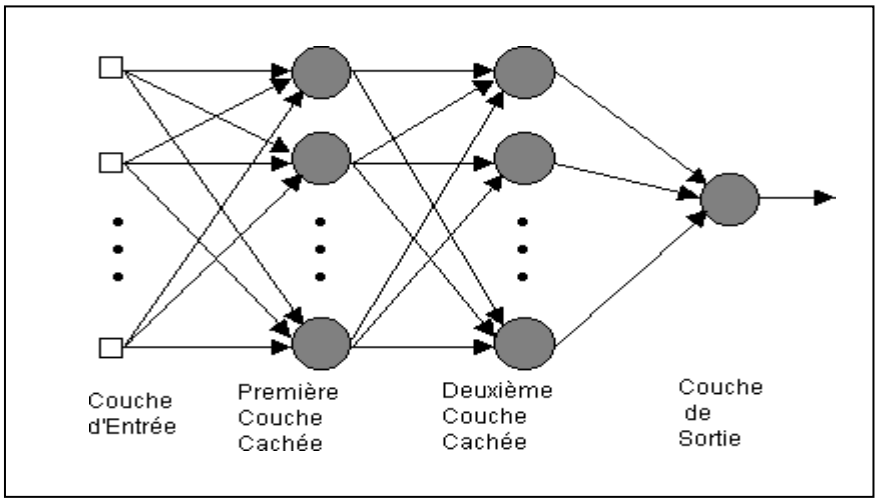

Figure I-3 : Réseau de neurones.

I-7-2 Les Arbres De Décision :

Les arbres de décisions sont des outils d'aide à la décision qui permettent selon des variables discriminantes de répartir une population d'individus en groupes homogènes en fonction d'un objectif connu. Les arbres de décision sont des outils puissants et populaires pour la classification et la prédiction. Un arbre de décision permet à partir des données connues sur le problème de donner des prédictions par réduction, niveau par niveau, du domaine des solutions.

Chaque nœud interne d'un arbre de décision permet de répartir les éléments à classifier de façon homogène entre ses différents fils en portant sur une variable discriminante de ces éléments. Les branches qui représentent les liaisons entre un nœud et ses fils sont les valeurs discriminantes de la variable du nœud. Et en fin, les feuilles d'un arbre de décision représentent les résultats de la prédiction des données à classifier [8]. Le **Tableau I-1** donne les Algorithmes D'induction Des Arbres De Décision

Nom de l'algorithme	*Développeur*	*Année*
CHAID	Kass	1980
CART	Breiman, et al.	1984
ID3	Quinlan	1986
C4.5	Quinlan	1993
SLIQ	Agrawal, et al.	1996
SPRINT	Agrawal, et al.	1996

Tableau I-1: Les algorithmes d'inductions des arbres de décision [7].

I-7-3 Les Algorithmes Génétiques

Un algorithme génétique se constitue d'une catégorie de programmes dont le principe est la reproduction des mécanismes de la sélection naturelle pour résoudre un problème donné. L'optimisation des problèmes combinatoires et surtout les problèmes dits NP-complets (dont le temps de calcul croit de façon non polynomiale avec la complexité du problème) est l'objectif principale des algorithmes génétiques, ils sont particulièrement adaptés à ce type de problèmes. Ces algorithmes constituent parfois une alternative intéressante aux réseaux de neurones mais sont le plus souvent complémentaires.

Principe De Base Des Algorithmes Génétiques : Le principe de fonctionnement d'un algorithme génétique est le suivant :

1. Codage du problème sous forme d'une chaîne binaire.
2. Génération aléatoire d'une population. Celle-ci contient un pool génétique qui représente un ensemble de solutions possibles.
3. Calcul d'une valeur d'adaptation pour chaque individu. Elle sera fonction directe de la proximité des différents individus avec l'objectif, on parle ici d'évaluation (fitness).
4. Sélection des individus doit se reproduire en fonction de leurs parts respectives dans l'adaptation globale.
5. Croisement des génomes des parents.
6. Sur la base de ce nouveau pool génétique, on repart à partir du point ***3***.

Codage D'un Algorithme Génétique : Avant de pouvoir utiliser un algorithme génétique pour résoudre un problème, il faut trouver un moyen pour encoder une solution potentielle à ce problème (les chromosomes), le codage consiste alors à choisir les structures de données qui coderont les gènes. Il existe différentes manières de le faire :

- **Codage binaire :** ce type de codage se base sur le principe de coder la solution selon une chaîne de bits (0 ou 1).
- **Codage à caractères multiples :** les chromosomes d'un algorithme génétique peuvent être codés d'une autre manière qui est le codage à l'aide de caractères. Souvent, ce type de codage est plus naturel que son

précédent.

- **Codage sous forme d'arbre :** il utilise une structure arborescente avec une racine (parent) de laquelle peuvent être issus un ou plusieurs fils (descendants).

I-7-4 Les Règles Associatives

Les règles associatives sont des règles extraites d'une base de données transactionnelles et qui décrivent des associations entre certains éléments. Elles sont fréquemment utilisées dans le secteur de la distribution des produits où la principale application est *l'analyse du panier de la ménagère* (Market Basket Analysis) dont le principe est l'extraction d'associations entre produits sur les tickets de caisse. Le but de la méthode est l'étude de ce que les clients achètent pour obtenir des informations sur qui sont les clients et pourquoi ils font certains achats.

La méthode recherche quels produits tendent à être achetés ensemble. La méthode peut être appliquée à tout secteur d'activité pour lequel il est intéressant de rechercher des groupements potentiels de produits ou de services : services bancaires, services de télécommunications, par exemple. Elle peut être également utilisée dans le secteur médical pour la recherche de complications dues à des associations de médicaments ou à la recherche de fraudes en recherchant des associations inhabituelles.

Une *règle d'association* est de la forme : Si **condition** alors **résultat**. Dans la pratique, nous nous limitons généralement à des règles où la condition se présente sous la forme d'une conjonction d'apparition d'articles et le résultat se constitue d'un seul article. Par exemple, une règle à trois articles sera de la forme : Si X et Y alors Z ; règle dont la sémantique peu être énoncée : Si les articles X et Y apparaissent simultanément dans un achat alors l'article Z apparaît **[7] [8]**. Le **Tableau I-2** présente les algorithmes d'inductions des règles associatives.

Nom de l'algorithme	*Développeur*	*Année*
APRIORI	Agrawal, et al.	1993
FP-GROWTH	Han, et al.	2000
ECLAT	Zaki	2000
SSDM	Escovar, et al.	2005
KDCI	Orlando, et al.	2003

Tableau I-2 : Les algorithmes d'inductions des règles associatives [8].

I-7-5 L'algorithme Des K-Plus Proches Voisins

L'algorithme des k plus proches voisins (K-PPV, k nearest neighbor en anglais ou kNN) est un algorithme de raisonnement à partir de cas qui est dédié à la classification qui peut être étendu à des tâches d'estimation. Le but de cet algorithme est de prendre des décisions en se basant sur un ou plusieurs cas similaires déjà résolus en mémoire.

Dans ce cadre, et Contrairement aux autres méthodes de classification (arbres de décision, réseaux de neurones, algorithmes génétiques, …etc.) l'algorithme de KNN ne construit pas de modèle à partir d'un échantillon d'apprentissage, mais c'est l'échantillon d'apprentissage, la fonction de distance et la fonction de choix de la classe en fonction des classes des voisins les plus proches, qui constituent le modèle **[9] [10]**.

Algorithme de classification par k-PPV

Paramètre : le nombre k de voisins

Donnée : un échantillon de m exemples et leurs classes La classe d'un exemple X est c(X)

Entrée : un enregistrement Y

1. Déterminer les k plus proches exemples de Y en calculant les distances
2. Combiner les classes de ces k exemples en une classe c

Sortie : la classe de Y est c(Y)=c **[11]**.

Comment k-PPV marche-t-il ?

Nous supposons avoir une base de données d'apprentissage constituée de *N* couples « entrée-sortie ». Pour estimer la valeur de sortie d'une nouvelle entrée *x*, la méthode des *k* plus proches voisins consiste à prendre en compte (de façon identique) les *k* échantillons d'apprentissage dont l'entrée est la plus proche de la nouvelle entrée *x*, selon une distance à définir **[12]**.Si nous prenons une base d'apprentissage de 100 éléments, Dès que nous recevons un nouvel élément que nous souhaitons classifier, l'algorithme calcule sa distance à tous les éléments de la base. Si cette base comporte 100 éléments, alors il va calculer 100 distances et donc obtenir 100 nombres réels. Si k = 25 par exemple, il cherche alors les 25 plus petits nombres parmi ces 100 nombres qui correspondent donc aux 25 éléments de la base qui sont les plus proches de l'élément que nous souhaitons classifier. La classe attribuée à l'élément à classifier est la classe majoritaire parmi ces 25 éléments **[13]**.

I-7-6 L'algorithme Des K-Moyennes (K-Means)

L'algorithme des K-moyennes est dédié aux tâche de clustering, il permet de diviser une population donnée en K groupes homogènes appelés clusters. Le nombre de clusters K est déterminé par l'utilisateur selon ses attentes.

Principe de fonctionnement : Après avoir déterminé un nombre K de clusters nous positionnons les K premiers points (appelés graines) au hasard (nous utilisons en général les K premiers enregistrements). Chaque enregistrement est affecté à la graine dont il est plus proche (en utilisant la fonction de distance). A la fin de la première affectation, la valeur moyenne de chaque cluster est calculée et la graine prend cette nouvelle valeur. Le processus est répété jusqu'à stabilisation des clusters [14] [15].

Algorithme de clustering par *K-Means:* L'algorithme *k-means* est en 4 étapes :

1. Choisir k objets formant ainsi k clusters
2. (Ré) affecter chaque objet O au cluster C_i de centre M_i tel que dist(O,M_i) est minimal
3. Recalculer M_i de chaque cluster (le barycentre)
4. Aller à l'étape 2 s'il faut faire une affectation

Evaluation : Le but de la technique des k-means est le regroupement par similitude des populations statiques, ce qui nécessite de déterminer la qualité des différents clusters. Une évaluation de la qualité pourrait consister à l'étude de la variance de cette population. Alors, nous pouvons dire qu'un cluster possédant d'une part population et d'autre part une variance faible est un cluster solide. Il est nécessaire d'effectuer d'autres évaluations dans les cas suivants:

- Si la population d'un cluster est trop faible, il sera préférable de grouper ce cluster avec un autre.
- Si un cluster est trop dominant, il pourrait être valable de diviser la population en deux (dans et hors cluster) et de relancer le processus pour chaque sous groupe [**14**].

I-8 Les Tâches de la fouille de données

Beaucoup de problèmes intellectuels, économiques ou même commerciaux peuvent être exprimés en termes des tâches suivantes : **[16]**

- ✓ La classification.
- ✓ L'estimation.
- ✓ La prédiction
- ✓ Le groupement par similitude (règles d'association).
- ✓ L'analyse des clusters.
- ✓ La description.

Les trois premières tâches sont des exemples de la fouille supervisée de données dont le but est d'utiliser les données disponibles pour créer un modèle décrivant une variable particulière prise comme but en termes de ces données. Le groupement par similitude et l'analyse des clusters sont des tâches non-supervisées où le but est d'établir un certain rapport entre toutes La description appartient à ces deux catégories de tâche, elle est vue comme une tâche supervisée et non-supervisée en même temps.

I-8-1 Classification :

La classification est la tache la plus commune de la fouille de données qui semble être une tâche humaine primordiale. Afin de comprendre notre vie quotidienne, nous sommes constamment obligés à classer, catégoriser et évaluer. La classification consiste à étudier les caractéristiques d'un nouvel objet pour l'attribuer à une classe prédéfinie. Les objets à classifier sont généralement des enregistrements d'une base de données, la classification consiste à mettre à jours chaque enregistrement en déterminant la valeur d'un champ de classe.

Le fonctionnement de la classification se décompose en deux phases. La première étant la phase d'apprentissage. Dans cette phase, les approches de classification utilisent un jeu d'apprentissage dans lequel tous les objets sont déjà associés aux classes de références connues. L'algorithme de classification apprend du jeu d'apprentissage et construit un modèle. La seconde phase est la phase de classification proprement dite, dans laquelle le modèle appris est employé pour classifier de nouveaux objets. **[17]**

I-8-2 L'estimation :

L'estimation est similaire à la classification à part que la variable de sortie est numérique plutôt que catégorique. En fonction des autres champs de l'enregistrement l'estimation consiste à compléter une valeur manquante dans un champ particulier. Par exemple on cherche à estimer la lecture de tension systolique d'un patient dans un hôpital, en se basant sur l'âge du patient, son genre, son indice de masse corporelle et le niveau de sodium dans son sang. La relation entre la

tension systolique et les autres données vont fournir un modèle d'estimation. Et par la suite nous pouvons appliquer ce modèle dans d'autres cas. **[18]**

I-8-3 La Prédiction :

La prédiction est la même que la classification et l'estimation, à part que dans la prédiction les enregistrements sont classés suivant des critères (ou des valeurs) prédites (estimées). La principale raison qui différencie la prédiction de la classification et l'estimation est que dans la création du modèle prédictif on prend en charge la relation temporelle entre les variables d'entrée et les variables de sortie**]**. Quelques exemples de l'utilisation des tâches de prédiction dans les domaines de recherche et commerce sont les suivants :

- Prévoir le prix des actions dans les trois prochains mois
- Prévoir le champion de la coupe du monde en football en se basant sur la comparaison des statistiques des équipes Prévoir quels clients va déménager dans les 6 mois qui suivent.

I-8-4 Le Groupement Par Similitude (Analyse des associations et de motifs séquentiels) :

Le groupement par similitude consiste à déterminer quels attributs "vont ensemble". La tâche la plus répandue dans le monde du business, est celle appelée l'analyse d'affinité ou l'analyse du panier du marché, elle permet de rechercher des associations pour mesurer la relation entre deux ou plusieurs attributs. Les règles d'associations sont, généralement, de la forme "Si <antécédent>, alors<conséquent>".

I-8-5 L'analyse Des Clusters :

Le clustering (ou la segmentation) est le regroupement d'enregistrements ou des observations en classes d'objets similaires. Un cluster est une collection d'enregistrements similaires l'un à l'autre, et différents de ceux existants dans les autres clusters. La différence entre le clustering et la classification est que dans le clustering il n'y a pas de variables sortantes. La tâche de clustering ne classifie pas, n'estime pas, ne prévoit pas la valeur d'une variable sortantes. Au lieu de cela, les algorithmes de clustering visent à segmenter la totalité de données en des sous groupes relativement homogènes. Ils maximisent l'homogénéité à l'intérieur de chaque groupe et la minimisent entre les différents groupes.

I-8-6 La Description :

Parfois le but de la fouille est simplement de décrire se qui se passe sur une Base de Données compliquée en expliquant les relations existantes dans les données pour premier lieu comprendre le mieux possible les individus, les produit et les processus présents dans cette base.

Une bonne description d'un comportement implique souvent une bonne explication de celui-ci. Dans les sociétés Algériennes nous pouvons prendre comme exemple comment une simple description, "les femmes supportent le changement plus que les hommes", peut provoquer beaucoup d'intérêt et promouvoir les études de la part des journalistes, sociologues, économistes et les spécialistes en politiques.

I-9 Domaines d'application

Le succès des méthode de data mining a intégré cette nouvelle science dans tous les domaines d'intelligence artificielle. Les méthode de data mining sont applicables au problème d'estimation, prévention , analyse de risque, catégorisation , reconnaissance et etc. Nous citons ici quelques applications connues du data mining:

- **Multimédia** : les techniques du data mining sont appliquées aux bases de données multimédia pour la résolution de certains problème tel que la recherche d'image par le contenu, la reconnaissance de forme, la reconnaissance de la voie et etc.
- **Application de finance**: la fouille de données a fait ses premiers succès dans le domaine des finances. Des problèmes de prédiction et de prévention ont fait l'objet des méthodes statistiques qui ont donné naissance a` des algorithmes spécialises. Des méthodes pour l'analyse du comportement des clients, l'analyse des risques (assurances), l'estimation de rentabilité (banques) sont utilisées dans le secteur des finances.
- **Web Mining**:le web mining est l'analyse des données du web par les techniques du data mining. On distingue différent tâche du web mining: web content mining (texte, image,...), Web structure mining (liens hypertextes,...) et Web usage mining (analyse des fichiers logs client et serveur).
- **Texte Mining**: la fouille de texte est utilisé dans divers domaines, les moteurs de recherche, la reconnaissance de l'´ecriture imprimée ou manuscrite et etc...
- **Médecine** : la fouille de données s'applique aussi dans le domaine médicale tel que le diagnostic automatique ou l'aide au diagnostic (découverte de la maladie du patient d'après ses symptômes), recherche du médicament le plus appropries a` une maladie et etc.

I-10 Conclusion

Fouille de données (Data mining) consiste à découvrir des modèles dans de grands ensembles de données, Les outils du data mining peuvent prédire les futurs tendances et actions, permettant de prendre les bonnes décisions. C'est ce qui rend le data mining la

technologie la plus importantes. en utilisant des méthodes situées à l'intersection du machine Learning, des statistiques et des systèmes de base de données afin d'identifier les modèles futurs. Il s'agit d'une étape essentielle du processus de découverte des connaissances, dans lequel des méthodes intelligentes sont appliquées à une grande quantité de données (big data) historiques pour extraire des modèles de données.

La fouille de données est une branche très importante, elle offre plusieurs méthodes et Techniques pour la résolution de différents problèmes. La fouille de données offre plusieurs techniques pour la résolution des problèmes nous nous intéressons aux techniques appliquez sur les données massive (big data) seront détaillés dans les chapitres qui suivent.

Chapitre II
Big Data

II-1 Introduction

Depuis plus de cinq (05) décennies l'informatique s'est implanté au cœur de nos entreprises, nos hôpitaux, nos ministères, nos foyers …..Etc. Cette forte utilisation de l'informatique à engendré de grands volumes de données qui ne sont pas gérable par les logiciels et matériels classique. Prenons le cas d'entreprises de taille humaine comme Google et Microsoft, ces grandes filiales qui doivent avoir des milliards de données à conservés.

Un autre exemple est celui des entreprises de téléphonie qui ont de grands volumes de données sur les clients et les prestations qui leurs sont offertes. Cette perplexité dans la gestion de ces grands volumes de données a donné naissance au Big Data. finitions et les Framework etc.., qui vous permettront à comprendre le concept du Big Data.

II-2 Définition

Littéralement, Big Data signifie données massives ou méga données. C'est un ensemble d'entités de données hétérogènes en extensibilité permanente qui ne peuvent pas pris en charge par les systèmes de gestion de données classiques.

Big Data est aussi une architecture distribuée et scalable pour le traitement et le stockage de grands volumes de données. En effet, on crée environ 2,5 milliards de Giga octets de données tous les jours, émanant des différents domaines créés par les divers outils numériques : vidéos publiés, messages envoyés, signaux GPS, enregistrements transactionnels d'achats en ligne et bien d'autres encore. Ces volumes massifs de données sont baptisés Big Data. Les géants du Web, au premier rang comme Yahoo, Facebook, Amazon et Google, ont été les tous premiers à déployer ce type de technologie pour permettre à tout le monde d'accéder en temps réel à leurs bases de données géantes.

L'émergence du Big Data est considérée comme une nouvelle révolution industrielle semblable à la découverte de la vapeur, de l'électricité, du téléphone et de l'informatique. D'autres, qualifient ce phénomène comme étant le dernier épisode de la troisième révolution industrielle, dite celle de « l'information ». Cependant, aucune définition universelle ou précise ne peut qualifier le Big Data. Etant un concept polymorphe et complexe, son interprétation varie selon les communautés qui s'y intéressent en tant que fournisseur ou utilisateur de services. Le Big Data est aussi défini par rapport à la manière avec laquelle les grandes masses de données peuvent être traitées et exploitées de façon optimale.

L'émergence du Big Data est considérée comme une nouvelle révolution industrielle semblable à la découverte de la vapeur, de l'électricité, du téléphone et de l'informatique.

D'autres, qualifient ce phénomène comme étant le dernier épisode de la troisième révolution industrielle, dite celle de « l'information ».

Depuis l'apparition de cette terminologie, beaucoup de bruit a été rouspété autour du Big Data, mais selon le Gartner, ce concept regroupe une famille d'outils qui répondent à une triple problématique dite règle des 3V Il s'agit notamment d'un **V**olume d'informations considérable à traiter, une grande **V**ariété de données (venant de diverses sources, non-structurées, organisées, Open…), et un certain niveau de **V**élocité à atteindre, autrement dit vitesse des échanges ou fréquence de création, collecte et partage de ces données [19].

Une autre définition similaire dans [19] ou les auteurs qualifient les Big Data comme n'importe quel type de source de données qui a au moins trois caractéristiques communes :

- **V**olumes de données extrêmement volumineux ;
- **V**itesse de transmission extrêmement élevée ;
- Très grande **V**ariété de données.

La vérité est que le Big Data a bien trois significations et que les éditeurs n'en abordent qu'une seule à la fois. Il est important de connaître leur positionnement pour comprendre le principe.

II-3 Les Caractéristiques Du Big Data

II-3-1 Volume

Le Big Data est associé à un volume de données vertigineux, se situant actuellement entre quelques dizaines de téraoctets (1 To=2_{12} octets) et plusieurs pétaoctets (1 Po=2_{15} octets) en un seul jeu de données. Le volume correspond à la masse d'informations produite chaque seconde. Les entreprises issues de tous les secteurs d'activité gérant des données massives, se voient assujetties à trouver des techniques nécessaires et moyens capables pour gérer les volumes de données collectés chaque jour et d'une importance vitale pour leur survie.

II-3-2 Vélocité

La vélocité ou la vitesse d'échanges décrit la fréquence à laquelle les informations sont générées, capturées, stockées et partagées. Elle est aussi le traitement des flux continus de données. Les entreprises doivent appréhender la vitesse non seulement en termes de création de données, mais aussi sur le plan de leur traitement, de leur analyse et de leur restitution à l'utilisateur en respectant les exigences des applications en temps réel.

II-3-3 Variété

De plus en plus, le taux des données structurées manipulées dans des tables de bases de données relationnelles est en décroissance par rapport à l'expansion des types de données non structurées. Cela peut être des images, des vidéos, des messages, des voix, et bien d'autres encore. Aujourd'hui, on trouve plusieurs milliers de sources hétérogènes comme les capteurs d'informations aussi bien dans les trains, les automobiles, les avions ou les équipements électroménagers qui émettent une variété d'information classement des données de différents types comme des photos sur différents sites ou les messages échangés sur les réseaux sociaux, etc. Ce sont les différents éléments qui constituent la variété supportée par le Big Data. Alex Popescu [20] parle aussi d'un autre V de **V**ariabilité qui l'a défini par le format et le sens des données qui peut varier au fil du temps.

II-3-4 Valeur

Toutes ces données massives ressemblent à une mine d'or potentielle, mais comme dans une mine d'or, vous avez seulement peu d'or et beaucoup plus de tout le reste. Parmi les défis les très sensibles de cette technologie est comment donner du sens à ces données pour tirer celles qui sont les plus pertinentes pour d'éventuelles prises de décisions dans une entreprise ?

Comment une organisation traite-t-elle des quantités massives de manière significative ? La notion de Valeur correspond à l'intérêt qu'on puisse tirer de l'utilisation de cette technologie. Selon les experts du domaine, les entreprises qui ne s'intéressent pas sérieusement au contenu de leurs volumes de données hébergées risquent d'être pénalisées et dépassées. Big Data désigne à la fois les grands volumes de données et la difficulté à extraire de cette masse de données celles ayant suffisamment de valeur pour justifier leur analyse. Big Data offre un ensemble d'outils d'analyse de données qui peuvent servir à préserver un privilège concurrentiel.

II-3-5 Véracité

L'aptitude à juger la crédibilité et la fiabilité du nombre indéfini de données collectées qualifie la Véracité du Big Data. Il est difficile de justifier l'authenticité et l'exactitude des contenus des différents volumes et variétés de données manipulées comme dans les conversations dans les réseaux sociaux avec les abréviations, le langage familier, les coquilles, les hashtags. La vérité est que le Big Data, a bien trois significations et que les éditeurs n'en abordent qu'une à la fois. Il est important de connaître leur positionnement pour leur poser les bonnes questions.

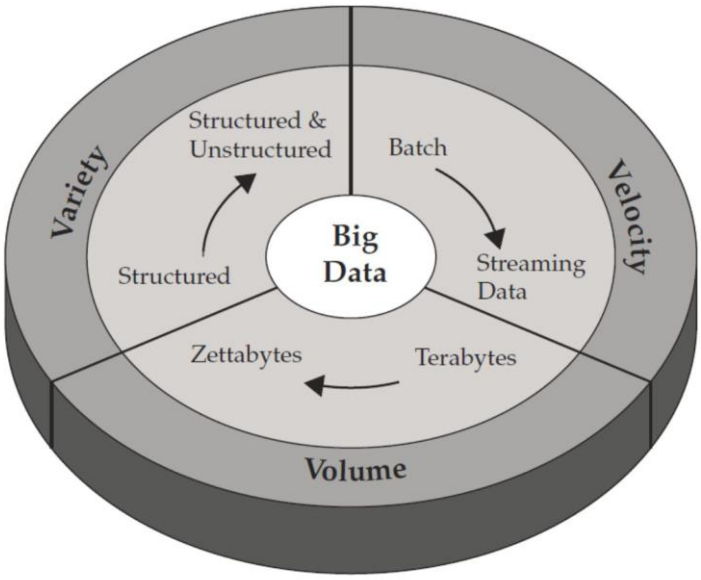

Figure II-1 Les caractéristiques du Big Data

II-4 Architecture Big Data

On distingue principalement les couches suivantes [19] :

- **Couche matériel (infrastructure Layer)**: Peut employer des serveurs virtuels VMware, ou des serveurs physiques ;
- **Couche stockage (Storage layer)**: Les données seront stockées soit dans une base NoSQL, ou bien directement dans le système de fichier distribué ou les Datawarehouses;
- **Couche management et traitement** : On trouve dans cette couche les outils de traitement et analyse des données comme MapReduce ou Pig ;
- **Couche visualisation** : pour la visualisation du résultat du traitement.

Une architecture Big Data doit inclure un ensemble de services qui permettent d'utiliser une multitude de sources de données de manière rapide et efficace. Parmi les composants de cette architecture, citons :

II-4-1 Infrastructure physique redondante

Les Big Data n'auraient probablement pas émergé, sans la disponibilité d'infrastructures physiques robustes différentes des infrastructures traditionnelles. L'infrastructure physique est basée sur une architecture distribuée ou les données peuvent être stockées physiquement dans plusieurs sites connectés par le biais des réseaux, un système de fichiers distribué et divers outils et applications d'analyse de données. La redondance est importante pour traiter des

données provenant de sources différentes, et pour assurer un service continu et réduire la latence.

II-4-2 Sécurité d'infrastructure

Plus les entreprises s'intéressent à l'analyse de leurs données contenues dans les Big Data, plus il sera important de sécuriser ces données. Un système Big Data doit authentifier les utilisateurs et prendre en chrge l'attribution de privilèges qui permet à ces utilisateurs de disposer légitimement d'un accès aux données. Ces types d'exigences de sécurité doivent faire partie de la conception préalable d'un système Big Data.

II-4-3 Avantages de l'architecture Big Data

Plusieurs avantages peuvent être associés à une architecture Big Data [19], citons par exemple :

- **Extensibilité (scalabilité)** : Le concept Big Data apporte une architecture scalable qui peut prévenir la taille d'infrastructure et l'espace disque nécessaire.

- **Performance** : Grâce au traitement parallèle des données et à son système de fichiers distribué, le concept Big Data est hautement performant en diminuant la latence des requêtes.

- **Coût faible** : On n'aura plus besoin de centraliser les données dans des baies de stockage souvent excessivement cher, grâce à unsystème de fichiers distribués, les disques internes des serveurs suffiront.

- **Disponibilité** : On a plus besoin des RAID disques, souvent coûteux. L'architecture Big Data apporte ses propres mécanismes de haute disponibilité.

II-5 Sources et types de données

Les données collectées, stockées et traitées dans les Big Data peuvent être issues de différents domaines et crées par plusieurs sources de données hétérogènes, ce qui génère une masse de données de types différents structurés et non structurés [19].

II-5-1 Sources de données structurées

Le terme « données structurées » désigne généralement des données dont la longueur et le format sont définis. Les exemples de données structurées comprennent des nombres, des dates et des chaînes (par exemple, le nom d'un employé, son poste de travail, etc.). On rappelle que la plupart des études déclarent que ce type de données représente environ de 10 à 20 pour cent

des données globales. Les données structurées sont les données qu'on a l'habitude de traiter, généralement stockées dans une base de données relationnelle ayant un schéma, interrogées à l'aide du langage de requête structuré (SQL). Elles sont recueillies à partir de sources traditionnelles.

Dans le monde des Big Data, les données structurées prennent un nouveau rôle avec l'évolution de la technologie qui offre de nouvelles sources de données structurées produites souvent en temps réel et en grands volumes. Les sources de données sont divisées en deux catégories :

- **Générées par ordinateur ou par machine** : Les données générées par machine se réfèrent généralement à des données créées par une machine sans intervention humaine.

- **Générées par l'homme :** Ce sont les données créées par les humains, en interaction avec les ordinateurs.

Dans la première catégorie, plusieurs sources existent actuellement :

- **Données de capteurs** : Données GPS, Radiofréquence, compteurs intelligents, dispositifs médicaux, journaux des appels en téléphonie, etc.

- **Données Web** : Le fonctionnement des serveurs, des applications et des réseaux permet de capturer toutes sortes de données comme les textes, les images, les vidéos, et tous ce qui peut figurer sur les pages Web.

- **Données commerciales** : Lorsque le caissier glisse le code-barres d'un produit acheté, toutes les données associées au produit sont générées. Ilsuffit de penser à tous les produits acquis par toutes les personnes, pour imaginer le volume de sonnées crée.

- **Données financières** : Beaucoup de systèmes financiers sont automatisés ; Ils sont exploités sur la base de règles prédéfinies qui automatisent les processus. Voici quelques exemples de données structurées générées par l'homme :

- **Données de saisie** : Il s'agit de toute donnée, qu'un être humain peut introduire dans un ordinateur, comme le nom, l'âge, le salaire, les sondages et ainsi de suite. Ces données peuvent être analysées pour comprendre le comportement des clients.

- **Données de flux de clics** : les données sont générées chaque fois qu'on clique sur un lien sur un site Web. Ces données peuvent être analysées pour déterminer le comportement des clients et les modèles d'achat.

- **Données liées aux jeux** : Chaque mouvement fait dans un jeu peut être enregistré. Cela peut être utile pour comprendre le comportement des utilisateurs.

II-5-2 Sources de données non structurées

Les données non structurées sont des données qui n'ont pas un format spécifié. Elles représentent 80 à 90 pour cent de l'ensemble des données disponibles. Jusqu'à quelques années auparavant, les outils disponibles n'offraient pas des traitements particuliers à ce type de données, mise à part le stockage ou l'analyse manuelle. Des données non structurées sont un peu partout : *pdf*, *doc*, email ou post dans un réseau social.

En fait, un grand nombre d'individus et organisations mènent leurs activités, leurs revenus autour de ces types de données. Tout comme pour les données structurées, les données non structurées sont générées soit par la machine ou par l'humain. Voici quelques exemples de données non structurées générées par une machine :

- **Images satellites** : Incluent les données météorologiques ou les données images de surveillance par satellite capturées par les services gouvernementaux. Un exemple typique de ces systèmes est Google Earth.

- **Données scientifiques** : Cela comprend l'imagerie sismique, les données atmosphériques, les données astronomiques, l'environnement, la génomique, la physique subatomique et la physique des hautes énergies[32].

- **Photographies et vidéo** : Concerne la sécurité, la surveillance et la vidéo de trafic. Voici d'autres exemples de données non structurées générées par l'homme :

- **Textes et courrier interne d'une entreprise** : Inclue tous les textes contenus dans les documents, les journaux, les rapports, les procès verbaux, les bilans, les résultats d'enquêtes, les sondages et les courriers. L'information d'entreprise représente en fait un grand pourcentage des données textuelles dans le monde d'aujourd'hui.

- **Données des médias sociaux** : Ces données sont générées à partir des plateformes des réseaux sociaux tels que YouTube, Facebook, Twitter, LinkedIn et Flickr.

- **Données mobiles** : Cela inclut des données telles que les messages texte et les informations de localisation.

- **Contenu du site Web** : Ceci provient de n'importe quel site offrant des contenus non structurés, tels que YouTube, Flickr ou Instagram [19,21].

Notons à la fin de cette section qu'il y a une autre catégorie de données qualifiée comme semi-structurée qui se place entre les deux catégories structurée et non structurée. Les données semi-structurées ne sont pas nécessairement conformes à une structure fixe prédéfinie mais peuvent être auto-descriptives et définies par des simples couples marque/valeur. Par exemple, ces couples peuvent inclure :<Famille> = Matallah, <Père> = Abdelkader, et <fils> = Oussama. Des exemples de données semi-structurées incluent XML (Extensible Markup Language), CSV file (Comma-Separated Values), EDI (Électronique Data Inter-change) et SWIFT (Langage de script implicitement parallèle).

II-6 Enjeux de Big Data

II-6-1 Enjeux Techniques :

Il existe essentiellement trois types de défis techniques autour du big data :

- Le stockage et la gestion des données massives, de l'ordre de la centaine de téraoctets ou du pétaoctet, qui dépassent les limites courantes des bases de données relationnelles classiques du point de vue du stockage et de la gestion des données.

- La gestion des données non-structurées (qui constituent souvent l'essentiel des données dans les scénarios Big Data), c'est-à-dire comment organiser du texte, des vidéos, des images, etc…

- L'analyse de ces données massives, à la fois pour le reporting et la modélisation prédictive avancée, mais également pour le déploiement.

II-6-2 Enjeux économiques :

D'après le cabinet de conseil dans le marketing IDC, « le marché du Big Data représentera 24 milliards de dollars en 2016, avec une part de stockage estimée à 1/3 de ce montant ». Il va sans dire que la « donnée » est le nouvel or noir du siècle présent, les spécialistes s'accordent déjà sur le fait que le Big Data sera l'arme économique de demain pour les entreprises et se présentera comme un levier qui fera la différence. Les entreprises collectent de plus en plus d'information en relation avec leurs activités (production, stockage, logistique, ventes, clients, fournisseurs, partenaires, etc), toutes ces informations peuvent être stockées et exploitées pour stimuler leur croissance.

II-6-3 Enjeux juridiques :

Le principal enjeu juridique reste la protection de la vie privée.

II-7 Conclusion

Le Big Data, la gestion des grands volumes de données à un champ d'application très vaste et varié. Dans un futur proche le Big Data serait très utile dans la création de nouvelles entreprises, de l'amélioration de la satisfaction clients, la détection d'épidémie. Lorsque nous avons une grande quantité de données, une question de base est de classifier les données en plusieurs groupes ou clusters. Par exemple, on peut classifier des articles de recherche par le thème ou des photos par le nom des personnes représentées. Comment effectuer cette tâche de regroupement de manière automatique ?

A ce niveau, ce n'est pas encore un problème algorithmique vu que l'objectif n'est pas clairement défini. Il faut d'abord modéliser les données. Une fois que nous avons fixé la représentation des données, il faut essayer de préciser l'objectif qu'on essaie d'atteindre de manière quantitative (et ceci peut dépendre du type de données auxquelles on a affaire). Ensuite, on cherche la technique pour optimiser cet objectif.

Chapitre III Techniques de Fouille de données sur Big Data

III-1 Introduction

Les dix dernières années ont connu une modification importante du volume et de la nature des données auxquelles le statisticien est confronté, la statistique est un élément important dans les Big Data parce que de nombreuses méthodes statistiques sont utilisées pour l'analyse des données massives.

L'évolution rapide des systèmes d'information génère des données de plus en plus volumineuses à causer de profonds changements de paradigme dans le travail de statisticien.

Les techniques statistiques classiques fonctionnent souvent bien avec des volumes de données moins importants, cependant dès que le volume de données devient massif, il y a un certain nombre de problèmes qui apparaissent.

III-2 Techniques du clustering (Classification supervisée)

III-2-1 Introduction

Le clustering (ou regroupement) est un sujet de recherche en apprentissage émanant d'une problématique plus générale, à savoir la classification. On distingue la classification supervisée et non supervisée. Dans le premier cas, il s'agit d'apprendre à classer un nouvel individu parmi un ensemble de classes prédéfinies, à partir de données d'entrainement (couples (individu, classe)). Issue des statistiques, et plus précisément de l'Analyse De Données (ADD), la classification non-supervisée, comme son nom l'indique, consiste à apprendre sans superviseur. A partir d'une population, il s'agit d'extraire des classes ou groupes d'individus présentant des caractéristiques communes, le nombre et la définition des classes n'étant pas donnés a priori. [22]

Le clustering (ou la classification non supervisée) est une approche importante en analyse exploratoire de données non étiquetées, mais reste un problème difficile. Sans connaissances a priori sur la structure d'une base de données, seule la classification non supervisée permet de détecter automatiquement la présence de sous-groupes pertinents (ou *clusters*).

Un cluster peut être défini comme un ensemble de données similaires entre elles et peu similaires avec les données appartenant à un autre cluster (homogénéité interne et séparation externe). Les clusters peuvent aussi être décrits comme des régions de l'espace de représentation des données contenant une densité relativement élevée de points de données, séparées entre elles par une zone de densité relativement faible. La détection de ces Regroupements joue un rôle indispensable pour la compréhension de phénomènes variés décrits par un ensemble d'observations. [23]

La classification non-supervisée fut, dans un premier temps, utilisée dans les domaines d'application suivants :

- la médecine : découverte de classes de patients présentant des caractéristiques physiologiques communes,
- le traitement de la parole : construction de systèmes de reconnaissance de l'orateur,
- l'archéologie : regroupement d'objets datant de l'âge de fer fia partir de leur description,
- l'éducation : structuration d'une classe d'ouvriers dans le domaine de l'industrie du téléphone, à partir de leurs besoins communs de formation.

La classification non-supervisée a trouvé différentes utilisations dans des domaines variés comme : la recherche et l'extraction d'information, dans des applications sur des données spatiales, par exemple, SIG (Système d'Information Géographique) ou sur des données astronomiques, l'analyse des données hétérogènes et séquentielles, les applications Web, l'analyse ADN en bioinformatique, etc. [24] (voir Tableau III.2)

Domaine	Formes de données	Clusters
Text mining	Textes Mails	Textes proches Dossiers automatiques
Web mining	Textes et images	Pages web proches
BioInformatique	Gènes	Gènes ressemblants
Marketing	Infos clients, produits achetés	Segmentation de la clientèle
Segmentation d'images	Images	Zones homogènes dans l'image

Tableau III.1 : Quelques domaines d'application du clustering

Des méthodes simples datant des années 1950-60 et issues des statistiques, telles que les approches par construction de taxonomies, offraient aux experts la possibilité d'analyser et d'explorer leurs données. Progressivement, d'autres applications sont apparues (traitement d'images, de données textuelles, etc.) imposant alors de nouvelles contraintes de performances, d'utilisations ou de flexibilités pour ces méthodes. Par ailleurs, le développement d'outils informatiques puissants a permis d'envisager de nouvelles pistes de recherche dans le domaine du clustering.

Depuis les années 1990, on peut considérer que le clustering constitue un domaine d'étude incontournable en apprentissage. Ces techniques occupent en effet une large place dans les processus de prétraitement, structuration, organisation des données et d'Extraction de Connaissances dans les Données (ECD). [22]

De nombreuses méthodes de classifications ont été proposées. Les approches les plus classiques sont les méthodes hiérarchiques et les méthodes partitives.

III-2-2 Concepts fondamentaux

III-2-2-1 Définition De La Classification :

Le concept de classification et la notion de partition d'un ensemble fini sont étroitement liés. Plusieurs techniques ont été définies se distinguant ou par le type de résultats obtenus ou alors par la méthode de regroupement qui constitue les classes (celles-ci se basent sur le calcul des distances).

Dans le premier cas selon qu'il y ait chevauchement entre les classes ou pas, on parlera de classification dure, douce ou floue ; selon qu'il y ait objet classé ou non on parlera de classification partielle. Ces mêmes résultats peuvent par ailleurs être représentés sous forme de structure plate ou d'une hiérarchie de classes. [25]

III-2-2-2 Partition d'un ensemble fini

Étant donné un ensemble fini d'objets Ω, on appelle partition de Ω toute famille de parties non vides de Ω disjointes dont l'union forme "ensemble. Ainsi, si C est une partition de Ω , alors :

$$C = \left\{ C_i \in P(\Omega) \,/\, \bigcap_{i=1}^{K} C_i = \{\Phi\}, \bigcup_{i=1}^{K} C_i = \Omega \right\}$$

Cette première définition correspond à l'algorithme de classification dure, c'est la manière la plus simple de représenter les résultats d'une classification ou chaque objet doit appartenir à une classe unique, c'est ce que l'on appelle méthodes de partitionnement. La Figure 1 représente un exemple de cette méthode.

a) Classification en 4classes *b) classification en 3 classes*

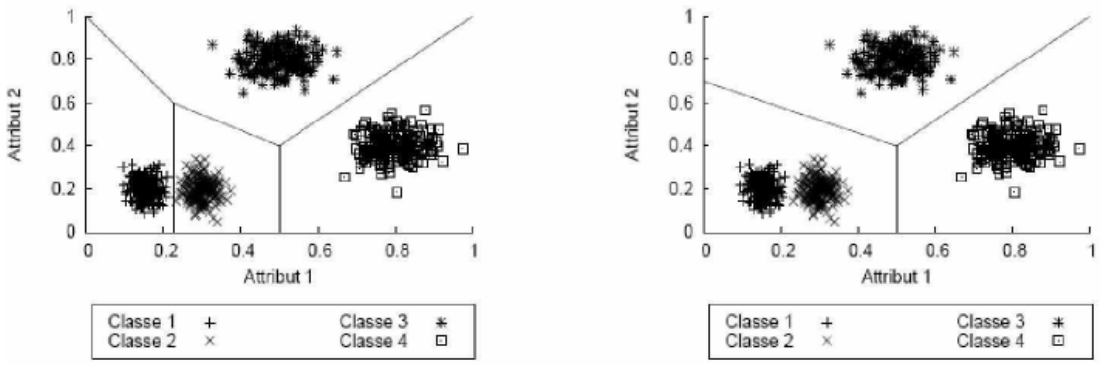

Figure III-1 Classification par partitionnement

Certes ce type de classification a l'avantage d'être aisément interprétable. Mais il est parfois inadéquat quand on a besoin de donner plus de flexibilité à la notion de classe afin de ne pas altérer la

qualité de l'information. Dans le cas où des objets sont trop différents des autres, il est préférable de ne pas les classer. Ainsi il s'agira de classification dure partielle.

Si la contrainte de disjonction est relâchée, un objet peut appartenir à différentes classes, on parlera de classification douce. Mais aussi de classification douce partielle s'il existe des objets qui n'appartiennent à aucune classe.

Enfin, dans le cas où chaque objet appartient à toutes les classes avec un degré d'appartenance, on va vers la notion de *classification floue.* Elle aboutit souvent vers une classification dure après une defuzzification des résultats par maximum d'appartenance par exemple, ou encore vers une classification douce ou les classes sont ordonnées par ordre décroissant du degré d'appartenance. Chaque objet est alors affecté aux classes dont il vérifie les conditions. [25]

III-2-2-3 Groupe D'objets Similaires

La similarité entre deux éléments i et j est noté $s(i,j)$. La similarité entre les données à étudier est l'information principale (et souvent unique) permettant à l'algorithme de classification de partitionner le jeu de donnée. Le plus souvent, cette similarité est calculée selon une mesure de "distance" adaptée au type de données et au problème à résoudre. Une distance, noté de manière générale $d(i,j)$, est une mesure de similarité qui vérifie certaines propriétés mathématiques (symétrie, séparation et inégalité triangulaire). Lorsque les données sont décrites dans un espace vectoriel, la distance la plus utilisée est la distance Euclidienne, alors notée$\|i - j\|$.

III-2-2-4 Les Fonctions Distance

Il est souvent utile d'avoir une représentation géométrique afin de fournir une première vue des données dans le but de repérer des groupes d'objets naturels ou des objets isolés. Chaque attribut représente alors un axe de coordonnées. L'espace des données est donc un espace euclidien à M dimensions dans lequel chaque objet est représenté par un point.

La distance $d(i, j)$ entre les points xi et xj, est ensuite mesurée dans cet espace grâce à l'utilisation d'une fonction de distance $d : \mathrm{R}^M \times \mathrm{R}^M$ dans R^+, respectant les propriétés suivantes [26]:

- $d(i, j) \geq 0$ (contrainte de positivité),
- $d(i, j) = 0$ si et seulement si $x_i = x_j$,
- $\forall x_i, x_j, d(i, j) = d(j, i)$ (contrainte de symétrie),
- $\forall x_i, x_j, x_l , d(i, j) \leq d(i, l)+d(l, j)$ (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 : il est très probable que des points proches dans l'espace de projection représentent des données d'un même groupe (la distance est alors proche de 0), alors que des points éloignés représentent, quant à eux, des données appartenant

à des groupes différents (la distance tend vers l'infini). La liste suivante non exhaustive, reprend les plus classiques :

- **La distance de Minkowski** : avec *p* un entier positif non nul.

$$d(i,j) = \sqrt[p]{\sum_{r=1}^{M} |x_{ir} - x_{jr}|^p}$$

- **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 $p = 2$:

$$d(i,j) = \sqrt{\sum_{r=1}^{M} |x_{ir} - x_{jr}|^2}$$

- **La distance de Manhattan** est également un cas particulier de la distance de Minkowski pour $p = 1$ (également appelée "métrique absolue") :

$$d(i,j) = \sum_{r=1}^{M} |x_{ir} - x_{jr}|$$

La notion de similarité est un élément essentiel de la classification automatique et le biais introduit par la mesure de similarité permet de former des groupes. Ce concept est dual à celui de la dis-similarité où deux individus sont autant plus similaires qu'ils sont proches au sens d'une mesure de dis-similarité.

Mesure de dis-similarité: *On appelle indice ou mesure de dis-similarité sur un ensemble* Ω*, une application* $d : \Omega \times \Omega \to R^+$ *qui vérifie les propriétés suivantes pour tout couple* $(x,y) \in \Omega \times \Omega$:

$$d(x,y) = d(y,x) \qquad \textit{(symétrie)}$$
$$d(x,y) = 0 \quad \Leftrightarrow \quad x = y \qquad \textit{(séparabilité)}$$

Métrique: *une métrique sur un ensemble* Ω *est une application* $d : \Omega \times \Omega \to R^+$ *qui vérifie les propriétés suivantes pour tout* $(x,y,z) \in \Omega \times \Omega \times \Omega$ *:*

$$d(x,y) = d(y,x) \qquad \textit{(symétrie)}$$
$$d(x,y) = 0 \quad \Leftrightarrow \quad x = y \qquad \textit{(séparabilité)}$$
$$d(x,y) \le d(x,z) + d(z,y) \qquad \textit{(inégalité triangulaire)}$$

Ultra-métrique: *On appelle ultra-métrique sur un ensemble, une application* $d : \Omega \times \Omega \to R^+$ *qui vérifie les propriétés suivantes pour tout* $(x,y,z) \in \Omega \times \Omega \times \Omega$ *:*

$$d(x,y) = d(y,x) \qquad \textit{(symétrie)}$$
$$d(x,y) = 0 \quad \Leftrightarrow \quad x = y \qquad \textit{(séparabilité)}$$
$$d(x,y) \le \max\{d(x,z), d(z,y)\} \qquad \textit{(inégalité ultramértique)}$$

L'homogénéité des individus regroupés au sein d'un groupe est souvent évaluée à l'aide d'un critère statistique appelée **variance** dont la définition est rappelée ci-dessous.

Variance: *On définit la variance $V(C_i)$ d un groupe d'objets C_i ainsi :*

$$V(C_i) = \frac{1}{N_i} \sum_{x_j \in C_i} d^2(x_j - \mu_i)$$

où N_i et μ_i sont respectivement le nombre d'objets et le centroïde du groupe C_i.

Dans la classification non-supervisé, on peut également distinguer la variance intra-classe V_{intra}, que l'on souhaite minimiser, de la variance inter-classe V_{inter}, que l'on cherche à maximiser :

$$V_{intra} = \frac{1}{N} \sum_{C_i \in C} N_i \times V(C_i)$$

$$V_{inter} = \frac{1}{N} \sum_{C_i \in C} N_i \times (\mu_i - \mu)^2$$

où N_i et μ_i sont respectivement le nombre d'objets et le centroïde du groupe C_i, et de manière analogue, N et μ désignent respectivement le nombre d'objets et le centroïde de Ω. La première évalue l'homogénéïté moyenne des groupes d'une partition et la seconde permet de quantifier la différence entre les groupes. Le théorème de König-Huyghens permet de relier la variance intra-classe et inter-classe à la variance totale $V_{totale} = V(\Omega)$ *:*

$$V_{totale} = V_{intra} + V_{inter}$$

III-2-2-5 Les Fonctions Similarités

De nombreux algorithmes conventionnels de classification utilisent la seule notion de distance afin d'obtenir des classes (ou des groupes) de points. En effet, il semble cohérent de regrouper les points dont la distance les uns par rapport aux autres est faible, et de séparer les points les plus éloignés. Cependant, d'autres algorithmes, plus récents, lui préfèrent 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, plus aisées à interpréter. [26]

La similarité $w(xi, xj)$ (notée $wi\ j$) entre deux points xi et xj est calculée en fonction de la distance entre ces points. La fonction similarité utilisée w : $R^N \times R^N$ dans [0,1] respecte les propriétés suivantes :

- $w_{i\,j} \subset [0,1]$ (contraintes de normalisation),
- $w_{i\,j} = 1$ si et seulement si $x_i = x_j$,
- $w_{i\,j} = w_{ji}$ (contraintes de symétrie).

Contrairement à la métrique de distance, deux points sont considérés comme similaires si la valeur de la similarité $w_{i\,j}$ (ou w_{ji}) est proche de 1. En revanche, ils sont considérés comme très différents l'un de l'autre si la valeur de la similarité $w_{i\,j}$ (ou w_{ji}) est proche de 0. Un exemple simple de similarité est la fonction binaire : $w_{i\,j} \in \{0,1\}$.

De la même façon que pour la fonction distance, il existe, dans la littérature, plusieurs définitions différentes de métrique de similarités entre objets décrits par des attributs. Parmi celles-ci, nous choisissons de présenter ***la fonction cosinus***, elle permet de mesurer la similarité entre deux vecteurs normalisés (en fixant leur norme à 1), en calculant l'angle entre ces derniers. Plus l'angle entre ces vecteurs est faible, plus les objets associés sont similaires. Cette fonction de similarité est utilisée essentiellement en exploration et analyse de contenu de documents.

$$w_{ij} = |\cos(x_i, x_j)| = \frac{|x_i{}^T x_j|}{||x_i||.||x_j||}$$

III-3 Les Trois Principales Etapes Du Clustering

Le processus de clustering se divise en trois étapes majeures : (1) la préparation des données, (2) l'algorithme de clustering et (3) l'exploitation des résultats de l'algorithme. Nous discutons ici des problématiques générales liées à chacune de ces étapes. [22]

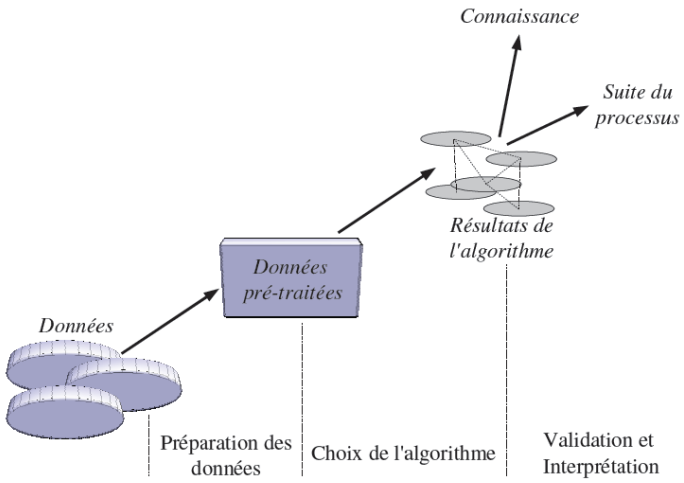

Figure III-2 Les différentes étapes du clustering

III-3-1 La préparation des données

- **Variables et sélection**

Les objets sont décrits par des variables, aussi appelées attributs, descripteurs ou traits. 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. En effet les variables ne sont pas nécessairement toutes pertinentes : certaines peuvent être redondantes et d'autres non-pertinentes pour la tâche ciblée. Ce problème de sélection de variables a été largement étudié en classification supervisée mais reste très ouvert pour une approche non-supervisée.

Dans un cadre supervisé, chaque variable peut-être évaluée relativement à son pouvoir discriminant par rapport aux classes prédéfinies. Deux types de méthodes se dégagent : les méthodes "*filtre*" (filter) et les méthodes "*enveloppe*" (wrapper). Dans le premier cas il
s'agit d'enlever les variables non pertinentes avant la phase d'apprentissage alors que les approches "enveloppe" utilisent de manière explicite le classifieur pour choisir un sous-ensemble de variables réellement discriminantes.

En revanche, peu de travaux concernent la sélection de variables dans une perspective non-supervisée, principalement parce que, sans les étiquettes de classe, il est difficile d'évaluer la pertinence d'une variable. Cette difficulté peut être contournée en effectuant une première étape de clustering à partir de l'ensemble des variables, puis de considérer chaque cluster comme une classe et de réitérer ce processus. De fait, cette technique se place parmi les approches "enveloppe" puisqu'elle est fortement dépendante de l'algorithme de clustering et des paramètres utilisés (nombre de clusters, etc.). Récemment, des approches "filtres" ont été envisagées dans ce contexte non-supervisé. [22]

- **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

donnés, 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. [28]

III-3-2 Le Choix De L'algorithme

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.)

- **La Taille Des Données**

La quantité d'objets à traiter est un premier facteur de décision. En effet, pour des données de très grande taille (par exemple en traitement d'images), les algorithmes de complexité plus que linéaires sont quasiment prohibés. Ainsi des méthodes telles que l'algorithme des $k\text{-}means$, proposé en 1967 par MacQueen ou plus généralement la méthode des nuées dynamiques, étant de complexité linéaire, sont très souvent utilisés. En revanche, lorsque l'on souhaite organiser quelques milliers, voire quelques centaines d'objets, il est possible d'avoir recours à des méthodes plus complexes et nécessitant un temps de traitement plus important (méthodes hiérarchiques ou de partitionnement plus élaborées). [22]

- **La Nature Des Données**

Beaucoup d'algorithmes de clustering s'appuient sur une matrice de similarité ou dis-similarité. Le plus souvent, cette matrice est obtenue à partir des descriptions des données. La nature de ces descriptions (variables qualitatives et/ou quantitatives), détermine alors le choix de la mesure de (dis) similarité utilisée.

Par ailleurs, on peut ne pas souhaiter traduire les données dans une telle matrice et conserver la table initiale des descriptions. Certaines méthodes, telles que le clustering conceptuel le permettent.

Quand bien même une matrice de similarité existe, certaines méthodes de clustering s'appuient sur la notion d'espace métrique. Pour l'algorithme des *k*-means par exemple, cet espace métrique permet de définir de nouveaux objets (ici les centroides) absents de l'ensemble initial des données. Dans le cas où un tel espace n'est pas présent, le processus de clustering se base uniquement sur la mesure de similarité. Des variantes sont alors envisageables telles que l'algorithme des nuées dynamiques proposé par Diday en 1972 ou plus

simplement les algorithmes "*k*-médoides" ou PAM (Partitioning Arround Medoids).

- **La Forme Des Clusters**

Les variations de taille et densité sont également à prendre en compte dans le choix de l'algorithme de clustering. On entend, par variation de taille, la capacité d'un algorithme à obtenir à la fois des clusters contenant beaucoup d'objets, et des clusters contenant peu voire très peu d'objets. De même, la prise en compte de la densité permet ou non d'obtenir des clusters contenant des objets plus ou moins "proches", au sens de la mesure de similarité établie. [22].

- **Le Type De Résultats Attendus**

La sortie d'un algorithme de clustering peut être, par exemple, une partition (ou pseudo-partition), une fonction ou encore un dendrogramme (arbre hiérarchique).

De même, chaque cluster obtenu peut être défini soit par l'ensemble des objets qui le composent, soit par une description relative aux variables initiales. On parle d'une définition en extension, dans le premier cas, et en intension, dans le second. Précisons que la plupart des approches proposent une classification à partir d'une définition en extension des clusters.

Une nouvelle fois, le choix de la méthode de clustering devra être fait en fonction du type de résultat souhaité et donc de l'exploitation envisagée de ce résultat.

III-3-3 L'exploitation Des Clusters

La tentation est grande, pour un non-spécialiste, de considérer comme "acquis" le résultat d'un processus de clustering. Autrement dit, 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 (méloïdés, 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). [22]

III-4 Les méthodes du clustering (classification non supervisée)

L'objectif de la classification automatique est de former des groupes d'individus ou de variables a fin de structurer un ensemble de données. On cherche souvent des groupes homogènes c'est à dire que les objets d'une même classe doivent être « similaires » et les objets de deux classes différentes doivent être «distincts». Les méthodes de classification se distinguent entre autre par la structure de classification obtenue (partition, hiérarchie…etc.). On distingue deux familles :

- Méthodes Hiérarchiques (classification hiérarchique, …)
- Méthodes à Partitionnement (k-means, …)

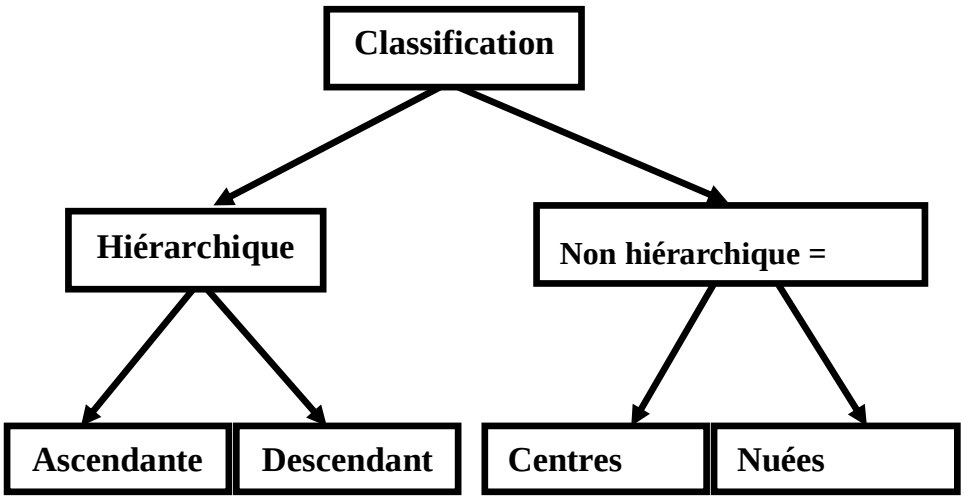

Figure III-3 Les méthodes de classification

Une multitude d'algorithmes de classification sont proposés dans la littérature qui peut être groupés selon différents critères :

- Le type de données utilisées pour l'approche
- Le critère de classification qui définit la similarité entre les observations
- La théorie et les concepts de base sur lesquels les techniques de classification sont basées (par exemple les statistiques, les réseaux de neurones, ...)

Ainsi par rapport à la méthode utilisée pour définir les classes, les algorithmes peuvent être classifiés de façon générale dans les catégories suivantes [27] :

- **La Classification Hiérarchique**. Le résultat de ce type d'algorithmes est un arbre de clusters, appelé dendogramme, qui montre comment les clusters sont organisés. En coupant le dendogramme au niveau désiré, une classification des données dans des groupes disjoints est ainsi obtenue.

- **La classification par partitionnement** cherche à décomposer directement la base de données dans un ensemble de clusters disjoints. Plus spécifiquement, ils essaient de déterminer un nombre de partitions qui optimisent certains critères (fonction objective). La fonction objective peut mettre en valeur la structure locale ou globale des données et son optimisation est une procédure itérative.
-

III-4-1 La Classification Hiérarchique

La classification hiérarchique construit une hiérarchie de clusters, ou autrement dit, un arbre de clusters, connue aussi sous le nom de dendogramme. Une telle approche permet d'explorer les données à travers différents niveaux de granularité. Les méthodes de la classification hiérarchique sont divisées en deux types d'approches : ascendante (CHA), dite agglomérative et descendante (CHD), dite divisive.

La première (ascendante) qui est la plus couramment utilisée consiste, à construire l'hiérarchie à partir des objets (au départ on a un objet par classe), puis à fusionner les classes les plus proches, afin de n'en obtenir plus qu'une seule contenant tous les objets. La seconde (descendante), moins utilisée, consiste à construire l'hiérarchie à partir d'une seule classe regroupant tous les objets, puis à partager celle-ci en deux groupes. Cette opération est répétée à chaque itération jusqu'à ce que toutes les classes soient réduites à des singletons. Le processus continu jusqu'à ce qu'un critère d'arrêt (d'habitude c'est le nombre *k* de clusters demandé) est atteint.

Les avantages de la classification hiérarchique incluent :

- Facilité pour traiter différentes formes de similarité ou de distance entre objets
- Applicable aux différents types d'attributs
- Une flexibilité en ce qui concerne le niveau de granularité

Les points faibles sont liés au :

- Choix du critère d'arrêt qui reste vague
- Fait que la plupart des algorithmes hiérarchiques ne revoient pas les clusters intermédiaires qu'une fois qu'ils sont construits pour les améliorer

Les méthodes de classification hiérarchique basées sur des métriques de liens donnent comme résultats des clusters de formes convexes.

Dans la classification hiérarchique la représentation des données utilisée d'habitude, *observation-variable*, n'est pas très significative. Au lieu de cela on utilise une matrice de distance$N \times N$, qui calcule les similarités ou dis-similarités entre les données.

Elle est parfois appelée matrice de *connectivité*. Les critères du lien sont calculés à partir de cette matrice. Parfois, par rapport à la mémoire machine, la taille de la matrice des distances est trop grande. Pour relaxer cette limite, différentes heuristiques sont utilisées pour introduire dans cette matrice une densité plus faible.

Cela peut être réalisé en omettant des entrées inférieures à un certain seuil, ou en utilisant seulement un certain sous-ensemble de représentants des données, ou en gardant pour chaque individu qu'un certain nombre des plus proches voisins.

Une matrice creuse peut-être utilisée pour représenter intuitivement les concepts de connectivité et de proximité. On remarque que la manière dont on traite la matrice de (dis) similarité et on construit une métrique de liens répète tout simplement nos idées a priori sur le modèle des données. [24].

L'algorithme général du clustering ascendant

Début

Entrées : n clusters de 1 donnée chacun

A - 1ère étape

• Calcul des distances entre chaque couple de clusters (n.(n-1)/2 distances)

• Identification de la distance d(Ci, Cj) minimale

• Regroupement des clusters Ci et Cj

B - Etapes suivantes

• Calcul des distances entre le nouveau cluster et les autres clusters (les autres distances restent inchangées)

• Identification de la distance minimale

• Regroupement des clusters associés

Si nbr de clusters > 1 : retour au point B

Sinon : fin

Fin

III-4-2 La classification non hiérarchique (par partitionnement)

Les algorithmes de partitionnement de données partagent les données en plusieurs sous-ensembles. Puisque l'examen de tous les sous-ensembles possibles est infaisable du point de vue computationnel, quelques heuristiques gloutonnes sont utilisées sous forme d'optimisation itérative. Plus précisément, cela correspond aux différents schémas de réallocation qui réaffectent itérativement les points entre les k clusters. Par rapport aux méthodes hiérarchiques traditionnelles,

dans lesquelles les clusters ne sont pas revus après avoir été construits, les méthodes par réaffectations améliorent les clusters progressivement.

Une des voies pour réaliser le partitionnement de données, c'est d'avoir un point de vue conceptuel qui identifie le cluster avec un certain modèle dont les paramètres inconnus doivent être estimés. Plus exactement, les modèles probabilistes supposent que les données proviennent d'un mélange de plusieurs populations dont on ne connait pas les distributions et d'autres paramètres à estimer. Un grand avantage des méthodes probabilistes est l'interpretabilité des classes obtenues.

En fonction de la méthode de calcul des prototypes, les méthodes de partitionnement par optimisation itérative sont divisées en k − médoïdes et k − moyennes (k − means). k − médoïde a deux avantages : le premier, est qu'il n'a pas de limites par rapport au type d'attributs et deuxièmement, le choix du medoïde est dicté par la location d'une fraction des points prédominants dans le cluster et ainsi est moins sensible à la présence des données aberrantes. Dans le cas des $k - moyennes$, un cluster est représenté par son centre, qui est une moyenne (d'habitude c'est une moyenne pondérée) des points affectés au cluster. Ce type d'algorithme est très adapté aux cas d'attributs numériques. D'une autre part, les centroïdes ont l'avantage d'avoir un sens géométrique et statistique assez clair. [24].

III-4-2-1 Les méthodes K-moyennes

L'algorithme $\boldsymbol{k - moyennes}$ ou $\boldsymbol{k - means}$ [28, 29] est l'outil de classification le plus utilisé dans les applications scientifiques et industrielles. Quoiqu'il ne fonctionne pas très bien pour les attributs catégoriels, il a un bon sens géométrique et statistique pour les attributs numériques. La somme des dispersions (distances) entre un point et son centroïde, exprimée par une distance appropriée, est utilisée comme fonction objective.

Par exemple, la fonction objective basée sur la norme L_2, la somme des carrées des erreurs (en Anglais Sum Of Squared Errors (SSE)) entre les points x_i et les centroïdes w_j correspondant, est égale à la variance intra-cluster totale :

$$E(C) = \sum_{j=1:K} \sum_{\mathbf{x}_i \in A} \|\mathbf{x}_i - \mathbf{w}_j\|^2$$

La somme des carrés des erreurs peut être rationalisée comme le log-vraisemblance pour les modèles de mélanges distribués normalement. Le but est de trouver un clustering avec une dispersion intra-cluster minimale. Par conséquent, *k-means* peut être dérivé du formalisme probabiliste général. On rappelle que seulement les moyennes sont estimées (dans le cas des *k-moyennes*). Une fonction objective basée sur la norme L_2 a plusieurs propriétés algébriques uniques.

Elle coïncide avec l'erreur par paire et avec la différence entre la variance totale des données et la variance interclasses :

$$E'(C) = \frac{1}{N} \sum_{j=1:K} \sum_{\mathbf{x}_i, \mathbf{y}_i \in A} \|\mathbf{x}_i - \mathbf{y}_i\|^2,$$

Par conséquent, la séparation des classes est atteinte en même temps que la consistance du cluster.

L'algorithme général de ces méthodes

Entrées : k le nombre maximum de classes désiré.

Début

1. Choisir k individus au hasard (comme centre des classes initiales)
2. Affecter chaque individu au centre le plus proche
3. Recalculer le centre de chacune de ces classes
4. Répéter l'étape (2) et (3) jusqu'à stabilité des centres
5. Éditer la partition obtenue.

Fin

Nous citons ici deux méthodes connues sur le principe de *k-means* sont :

- Méthodes de centres mobiles
- Méthodes des nuées dynamiques

➢ **La première version (*Méthodes de centres mobiles*)** consiste à réitérer les deux étapes suivantes : (1) réaffecter tous les points à leurs centroïdes proches, et (2) recalcule les centroïdes des nouveaux groupes formés. Les itérations continuent jusqu'à ce qu'un critère d'arrêt soit atteint (par exemple, le nombre d'itérations).

➢

Algorithme des méthodes de centres mobiles

Entrées : *k le nombre maximum de classes désiré.*

Début

1. Choisir k individus au hasard (comme centre des classes initiales)
2. Affecter chaque individus au centre le plus proche Ce qui donne une partition en k classes $P1 = \{C1, \ldots, C\text{k}\}$
3. On calcule les centres de gravité des chacune des classes de $P1$ Ce qui donne k nouveaux centres de classes.
4. Répéter l'étape (2) et (3) jusqu'à deux itérations successives donnent la même partition
5. Éditer la partition obtenue.

Fin

Cette version est connue comme l'algorithme de Forgy [30] et a plusieurs avantages :

- Il travaille facilement avec n'importe quelle norme L_p
- Il permet une simple parallélisation
- Il est insensible à l'ordre des données

➢ **La deuxième version *(Méthodes des nuées dynamiques)*** classique dans l'optimisation itérative, réaffecte des points basés sur l'analyse plus détaillée des effets sur la fonction objective causés par le mouvement d'un point de son groupe actuel vers un nouveau cluster potentiel.

Si le transfert a un effet positif, le point est déplacé et les deux centroïdes sont recalculés. Cette version est difficilement réalisable du point de vue computationnel, car elle nécessite une boucle propre pour chaque observation par déplacement des centroïdes.

Cependant, dans le cas L_2 tous les calculs peuvent-être algébriquement réduits à calculer simplement une seule distance. Par conséquent, dans ce cas les deux versions ont la même complexité de calcul. C'est une évidence expérimentale que la deuxième version (classique), comparée à algorithme de Forgy [30], fournit souvent de meilleurs résultats. En particulier, Dhillon et al. [31] ont remarqué que les k-moyennes sphériques de Forgy (utilisant une similarité basée sur le cosinus à la place d'une distance Euclidienne) ont tendance de bloquer quand ils sont appliqués à la classification de documents. Ils se sont aperçus qu'une version qui réaffecte les observations et recalcule immédiatement les centroïdes fonctionne beaucoup mieux. En plus de ces deux versions, il y a eu plusieurs tentatives pour trouver le minimum de la fonction de coût des k-moyennes.

La grande popularité des *k-means* est bien méritée. C'est un algorithme simple qui se base sur un fondement solide de l'analyse de la variance, mais il souffre des aspects suivants :

- Les résultats dépendent beaucoup de l'initialisation
- Le minimum local calculé semble être très loin du minimum global
- Le processus est sensible aux données atypiques.
- Les classes qui résultent peuvent être déséquilibrées (même vide, dans la version de Forgy [30])

III-5 Comparaisons entre K –Means et Clustering Hiérarchique [41]

Propriétés	K –Means	Clustering Hiérarchique
Définition	K signifie regroupement génère un nombre de disjoints, plat (non hiérarchique) Clusters	Classification hiérarchique méthode construire un hiérarchie du clustering, pas juste une seule partition de objets
Critères	Il est bien adapté à génération globulaire Grappe	Utilisez une matrice de distance comme Critères de clustering. La condition de résiliation peut être utilisé. Exemple –A nombre de clusters
Performance	La performance de L'algorithme K-moyenne est mieux que Hiérarchique Clustering Algorithme	Classification hiérarchique Les performances de l'algorithme sont moins par rapport à la moyenne K algorithme
Catégorie des données	K- Les moyens peuvent être utilisé dans catégorique les données sont les premières converti en numérique par attribuer le rang.	algorithme hiérarchique était adopté pour catégorique données, et en raison de son complexité une nouvelle approche pour attribuer une valeur de rang à chaque attribut catégoriel.
Sensiblilité au Bruit	K-Means est très sensible au bruit dans l'ensemble de données	Il est moins sensible au bruit dans l'ensemble de données.
Cluster	Il y a toujours K.	Le nombre de clusters k est pas nécessaire comme entrée.
Temps d'éxécution	Algorithme K -mean augmente également son temps d'exécution.	Algorithme hiérarchique son la performance est meilleure
Qualité	Algorithmes K-Means Montre moins de qualité	Algorithme hiérarchique montre plus de qualité
Base de données	L'algorithme k -mean est bon pour les grands base de données	Hiérarchique est bon pour petits ensembles de données

Tableau III-2 : Comparaisons entre K –Means et Clustering Hiérarchique

Chapitre IV Expérimentation et Résultats

VI-1 Fouille de données du Secteur de la Santé

Il existe plusieurs domaine d'utilisation des Big data, on a choisi le secteur de la santé pour appliquer les deux techniques de fouille de données.

L'analyse efficace et en temps réel des Big Data a déjà fait ses preuves dans le domaine de la santé. En effet, plusieurs modèles ont été testés pour améliorer le service médical privé et public, de même que la qualité de vie des patients, et ce, dans différents pays. Big Data peut encore révolutionner le domaine de la santé, non seulement en soutenant l'optimisation des services opérationnels, mais aussi en offrant des outils d'aide à la décision plus efficaces et en diminuant les coûts importants de ce secteur. En bref, l'exploitation et l'intégration adéquate de larges sources de données médicales apportent plusieurs opportunités notables, en particulier:

- L'optimisation des services et des dépenses médicaux : L'analyse du Big Data aide les organismes œuvrant dans le secteur de la santé à mieux détecter les services nécessitant une réorganisation et à suivre en temps réel la qualité des services rendus et la performance des unités médicales, de même que leurs besoins en approvisionnement humain et matériel.

- La personnalisation des services médicaux : A titre d'exemple, en exploitant l'analyse des données en temps réel, des modèles médicaux permettent de suivre à distance l'état des patients pour ajuster les doses ou faire des recommandations selon les symptômes relevés.

- Une meilleure prévention : Grâce à l'analyse avancée des flux de données cliniques collectés dans le secteur public et privé, les modèles prédictifs du Big Data peuvent aider à mieux planifier les moyens de prévention et à soutenir la gestion des épidémies, en particulier la détection précoce des signes alarmants touchant la santé de la population. Cela aide les décideurs à élaborer des plans de réponses optimisés selon le besoin de chaque région et selon la gravité des symptômes des individus.

- L'intégration de plusieurs sources médicales distribuées et hétérogènes constitue un défi de taille, afin de réussir ce pari et de mieux exploiter les opportunités du Big Data dans le secteur de la santé.

IV-2 Présentation du Domaine Etudié

C'est une enquête de satisfactions dans un hôpital, récupéré lors d'un cours de FUN (France Unité Numérique) [34]. Il s'agit d'une étude évaluant la qualité de relation et la quantité d'information reçue par le patient lors de son séjour à l'hôpital. 534 patients ont été recrutés sur plusieurs hôpitaux de la région parisienne.

IV-3 Description du jeu de données

Variables	description	Code/Unité
service	service ayant accueilli le patient	code de 1 à 8
sexe	sexe du patient	0 homme, 1 femme
age	âge	en années
profession	profession exercé par les patients	1 : agriculteur exploitant 2 : artisan, commerçant, chef d'entreprise 3 : cadre, profession intellectuelle ou artistique, profession libérale 4 : profession intermédiaire de l'enseignement, de la santé, du travail social ou de la fonction publique, 6 : ouvrier 7 : étudiant, militaire, chômeur sans avoir jamais travaillé 8 : autre
amelioration.sante	impression d'amélioration de la santé du fait du séjour à l'hôpital	0 : aggravée, à 3 : nettement améliorée
amelioration.moral	impression d'amélioration du moral du fait du séjour à l'hôpital	0 : aggravé, à 3 : nettement amélioré
recommander	recommander le service à son entourage	0 : non, 1 : oui, probablement, 2 : oui, sûrement.
score.information	score relatif à la qualité de l'information reçue pendant le séjour	score variant de 10 à 40
score.relation	score relatif à la qualité des relations avec le personnel soignant pendant le séjour	score variant de 10 à 40

Tableau IV-1 : Description du jeu de données

IV-4 Comment les algorithmes sont comparés ?

Les deux algorithmes de clustering sont comparés en fonction des facteurs suivants **[40.41.42]**:

- La taille de l'ensemble de données.
- Le nombre de clusters.
- Le type d'ensemble de données.
- Le type de logiciel.

Pour chaque facteur, quatre tests sont effectués, un pour chaque algorithme. Par exemple, selon la taille des données, chacun des quatre algorithmes: K-means, Hierarchical Clustering, est exécuté deux fois; d'abord en essayant un énorme jeu de données, puis en essayant un petit jeu de données. Le nombre total de fois que les algorithmes ont été exécutés est de 32. Pour chaque groupe de 8 essais, les résultats des exécutions sont étudiés et comparés. Les conclusions sont écrites. Cette étape est répétée pour tous les facteurs. En fonction du nombre de clusters, k (voir tableau IV- 2), sauf pour le clustering hiérarchique, tous les algorithmes de clustering comparés ici nécessitent de définir k à l'avance (k est le nombre de nœuds dans le réseau). Ici, les performances de différents algorithmes pour différents k sont comparées afin de tester les performances liées à k.

Pour simplifier la situation et faciliter les comparaisons, k est choisi égal à 8, 16, 32 et 64, Pour comparer le regroupement hiérarchique avec d'autres algorithmes, l'arbre hiérarchique est coupé à deux niveaux différents pour obtenir les nombres correspondants de clusters (8, 16, 32 et 64). En conséquence, à mesure que la valeur de k augmente, Cependant, les performances des algorithmes K-means meilleures que celles de l'algorithme de clustering hiérarchique.

Nombre des clusters K	**Performances**	
	k-Means	**CAH**
8	**63**	**65**
16	**71**	**74**
32	**84**	**87**
34	**89**	**92**

Tableau IV- 2 La relation entre le nombre de clusters et les performances de l'algorithme

- **Selon la précision** (voir tableau IV- 3) Mais à mesure que le nombre de k augmente, la précision de la classification hiérarchique devient meilleure jusqu'à ce qu'elle atteigne la précision. Les algorithmes K-means ont moins de qualité (précision) tous les algorithmes ont une certaine ambiguïté dans certaines données bruyantes pour être cluster.

Nombre des clusters K	**Qualité**	
	k-Means	**CAH**
8	**1112**	**1090**
16	**1089**	**960**
32	**910**	**850**
34	**840**	**760**

Tableau IV- 3 La relation entre le nombre de clusters et la qualité des algorithmes

- **Selon la taille du jeu de données** (voir tableau IV-4), un énorme jeu de données est utilisé, composé de 534 lignes et 60 colonnes et d'un petit jeu de données de 200 lignes et 20 colonnes. Le petit jeu de données est extrait en tant que sous-ensemble de l'énorme jeu de données. La qualité des algorithmes K-means devient très bonne lors de l'utilisation d'un énorme ensemble de données. L'algorithme de clustering hiérarchique montre de bon résultat lors de l'utilisation d'un petit ensemble de données. En conclusion, des algorithmes de partitionnement K-means sont utilisés pour un énorme ensemble de données tandis que des algorithmes de clustering hiérarchique sont utilisés pour de petites bases de données.

Taille des données	K=32	
	k-Means	CAH
36000	910	850
4000	95	91

Tableau IV-4 L'effet de la taille des données sur les algorithmes

- **Selon le type de jeu de données** (voir Tableau IV-5), un jeu de données aléatoire est utilisé, extrait d'Internet et utilisé pour différents emplois. D'autre part, un jeu de données idéal est utilisé qui fait partie du logiciel (LNKnet et Cluster et TreeView). Il est idéal car il est conçu pour être adapté pour tester et entraîner le logiciel lui-même et avoir des données moins bruyantes, ce qui conduit à l'ambiguïté. En conséquence, le clustering hiérarchique donne de meilleur résultat que l'algorithme K-means lors de l'utilisation d'un ensemble de données aléatoires et vice versa. Cela indique que l'algorithme K-means et très sensible au bruit dans l'ensemble de données. Ce bruit rend difficile pour l'algorithme d'inclure un objet dans un certain cluster. Cela affectera les résultats de l'algorithme.

Type de données	K=32	
	k-Means	CAH
Aléatoire	910	850
Idéale	810	829

Tableau IV-5 L'effet du type de données sur les algorithmes

- **Selon le type de logiciel**, deux packages sont utilisés pour comparer les algorithmes: LNKnet (environnement UNIX) et Cluster et TreeView (environnement WINDOWS).

Cependant, l'exécution des algorithmes de clustering en utilisant l'un d'entre eux donne presque les mêmes résultats même en modifiant l'un des trois autres facteurs (taille de l'ensemble de données, nombre de clusters et type de jeu de données). Ceci, est dû au fait que la plupart des logiciels utilisent les mêmes procédures et idées dans tout algorithme mis en œuvre par eux.

IV-5 Environnement de travail

IV-5-1 Environnement R

R est une suite intégrée d'équipements logiciels pour la manipulation de données, le calcul et l'affichage graphique. il comprend une installation de traitement des données efficace et stockage ,une série d'opérateurs pour les calculs sur les tableaux, dans des matrices particulières, une grande collection cohérente et intégrée d'outils intermédiaires pour l'analyse des données, installations graphiques pour l'analyse des données et l'affichage soit à l'écran ou sur papier, et un bien développé, langage de programmation simple et efficace qui comprend conditionnels, les boucles, les fonctions récursives définies par l'utilisateur et d'entrée et de sortie des installations.

Le terme « environnement » vise à caractériser comme un système entièrement planifié et cohérent, plutôt que d'une accumulation progressive d'outils très spécifiques et rigides, comme cela est souvent le cas avec d'autres logiciels d'analyse de données.

R est conçu autour d'un véritable langage informatique, et permet aux utilisateurs d'ajouter des fonctionnalités supplémentaires en définissant de nouvelles fonctions. Une grande partie du système lui-même est écrit dans le dialecte R, ce qui le rend facile pour les utilisateurs de suivre les choix algorithmiques faits. Pour les tâches informatiquement intensives, C, C ++ et Fortran peuvent être liés et appelé au moment de l'exécution. Les utilisateurs avancés peuvent écrire du code C pour manipuler des objets R directement.

De nombreux utilisateurs pensent de R en tant que système statistique. Nous préférons penser d'un environnement dans lequel sont mis en œuvre des techniques statistiques. R peut être étendu (facilement) par paquets. Il y a environ huit paquets fournis avec la distribution de R et beaucoup d'autres sont disponibles dans la famille CRAN de sites Internet couvrant un très large éventail de statistiques modernes.

IV-5-2 R Studio

R Studio est livré en deux versions : RStudio Desktop, pour une exécution locale du logiciel comme tout autre application, et RStudio Server qui, lancé sur un serveur Linux, permet d'accéder à

RStudio par un navigateur web. Des distributions de R Studio Desktop sont disponibles pour Microsoft Windows, OS X et GNU/Linux.

Depuis la version 1.0 (novembre 2016), RStudio intègre la possibilité d'écrire des notebooks combinant de manière interactive du code R, du texte mis en forme en markdown et des appels à du code Python ou Bash [33].

En devons spécifier le nombre de clusters dans lesquels nous voulons que les donnes soient regroupées. L'algorithme attribue de manière aléatoire chaque observation à un cluster et trouve le centroïde de chaque clustering. Préalablement à l'étape de partitionnement des données, il est indispensable de réaliser un prétraitement de la base, par l'Elimination ou traitement des données manquantes, codage, la transformation des variables qualitatives, réduction des données.

IV-6 Importation de la base de données

Nous avons sélectionné la base de données appelée " satisfaction-hôpital " pour le travail dans le projet et sont du type TXT contenant 11 colonne et élément 142311 **[34]**.

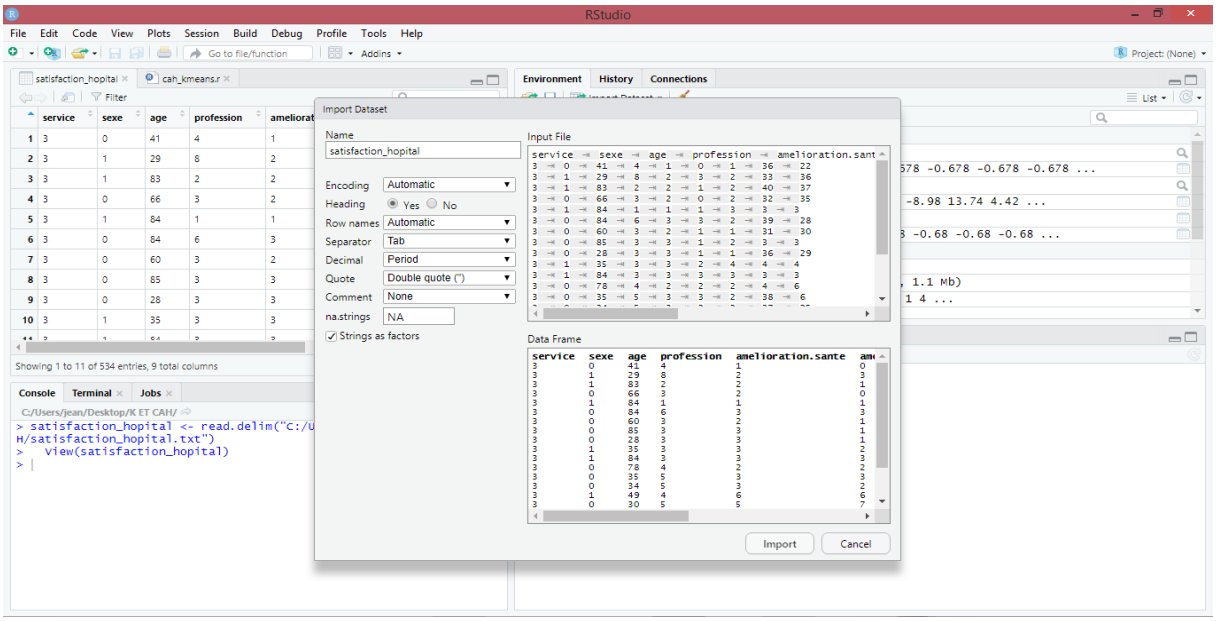

Figure IV-1: Import la base de données.

IV-7 Affichage de la base de données:

Pour l'affichage de base de données on utilise l'instruction suivant :

> View(satisfaction_hopital)

Le diagramme de paires est une matrice de diagrammes de dispersion qui est une visualisation très pratique pour balayer rapidement les corrélations entre de nombreuses variables dans un ensemble de données.

#graphique -croisement deux à deux pairs(satisfaction_ hôpital)

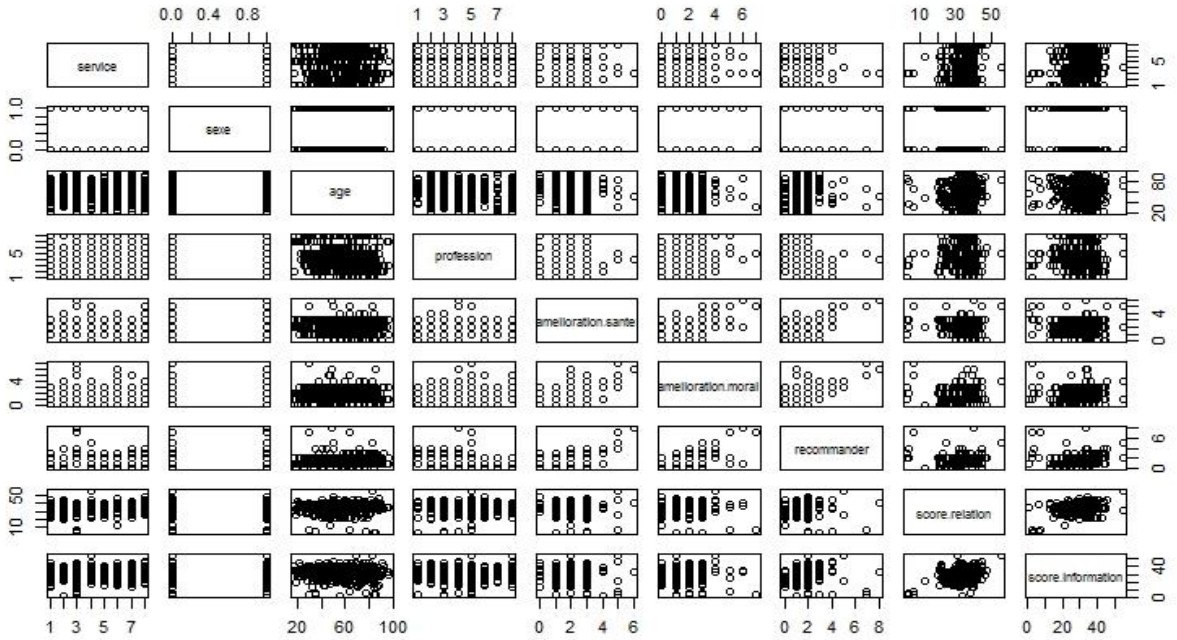

Figure IV-2: Croisement deux a deux.

IV-8 Exécution des algorithmes K-means et CAH

On peut effectuer le regroupement avec la fonction hclust(). Tout d'abord, calculons les valeurs de dissimilarité avec dist(), puis alimentons ces valeurs dans l'hclust et spécifions la méthode d'agglomération à utiliser (c'est-à-dire "complète", "average", "single", "ward.D"). Nous pouvons ensuite tracer le dendrogramme.

#CAH -critère de Ward**[35]**

#method= «ward.D2» correspond au vrai critère de Ward

> #affichage dendrogramme

cah.ward <- hclust(d.satisfaction_hopital,method="ward.D2")

> plot(cah.ward)

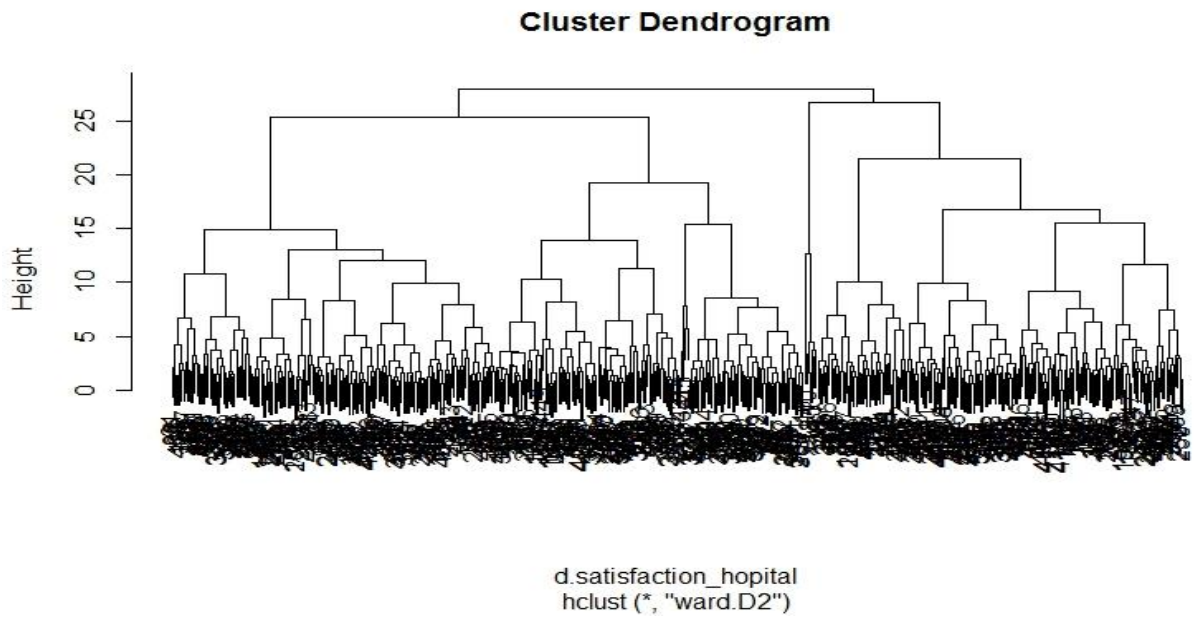

Figure IV-3: hclust dendrogramme

Dans les dendrogrammes, chaque feuille correspond à une observation. Au fur et à mesure qu'avançons dans l'arbre, des observations similaires sont combinés en branches, qui sont elles-mêmes fusionnés à une hauteur plus élevé.

La hauteur de la fusion, fournie sur l'axe vertical, indique la (dis) similitude entre deux observations. Plus la hauteur de fusion est élevée, moins les observations sont similaires. Il est à noter que les conclusions sur la similarité de deux observations ne peuvent être établies qu'en fonction de la hauteur ou les branches contenant ces deux observations sont fusionnées au préalable. Nous ne pouvons utiliser la proximité de deux observations le long de l'axe horizontal comme un critère de leur similitude.
La hauteur de la coupe au dendrogramme contrôle le nombre de clusters obtenus. Il joue le même rôle que k dans l'algorithme des K-moyennes. Afin d'identifier les sous-groupes (c.-à-d. les clusters)0

 # dendrogramme avec matérialisation des groupes[35]

> rect.hclust(cah.ward,k=4)

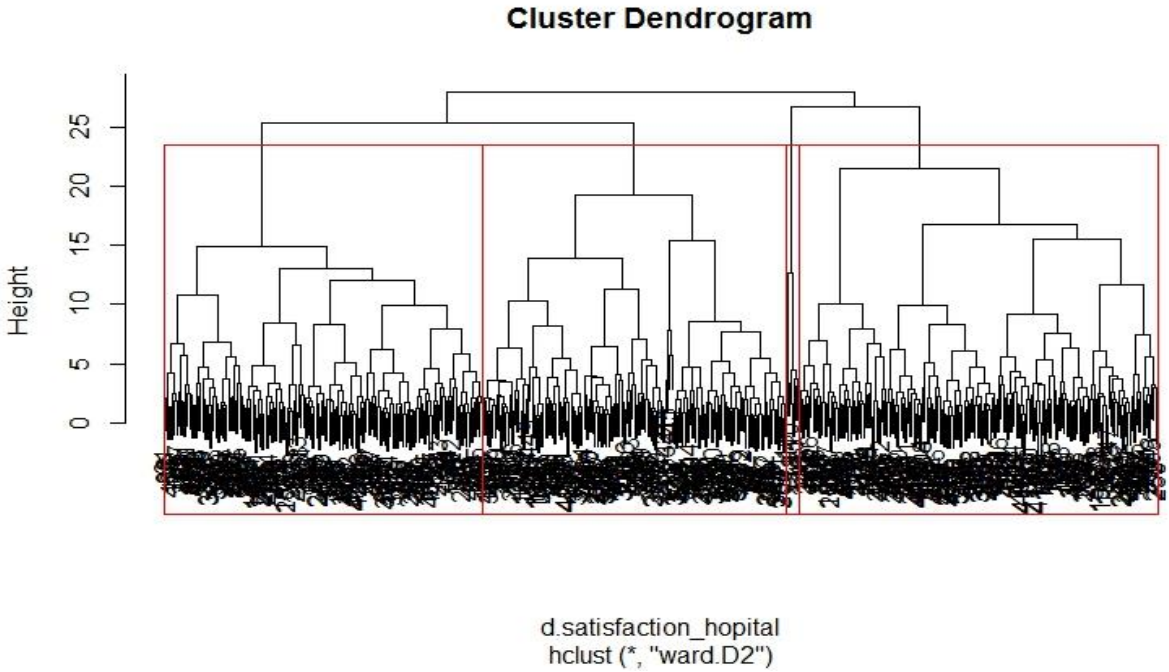

Figure IV-5: hclust dendrogramme avec matérialisation des groupes

Le regroupement peut être un outil très utile pour l'analyse des données dans le cadre non supervisé. Cependant, il existe un certain nombre de problèmes qui se posent lors de la mise en forme des clusters. Dans le cas du regroupement hiérarchique.

Avant d'aborder l'implémentation de l'algorithme des k-moyennes sous R, nous devons répondre à une question fondamentale qui est « comment choisir le nombre de clusters (K) ? » Le principe du partitionnement des données consiste à faire en sorte que les groupes soient regroupés de manière homogènes dans les clusters et de maniéré distincte des autres groupes. Il n'existe aucune formule mathématique qui peut nous donner directement une réponse au choix de "K", mais c'est un processus itératif ou nous devons exécuter plusieurs itérations avec différentes valeurs de "K" et choisir celles qui répondent le mieux à notre objectif

 # Choix Nombre of Clusters[[36.37.38.39]]

```
wss <- (nrow(newdata)-1)*sum(apply(newdata,2,var))
> for (i in 2:20)
+  wss[i] <- sum(kmeans(newdata,centers=i)$withinss)
> plot(1:20, wss, type="b", xlab="Number of Clusters",
```

+ ylab="Within groups sum of squares")

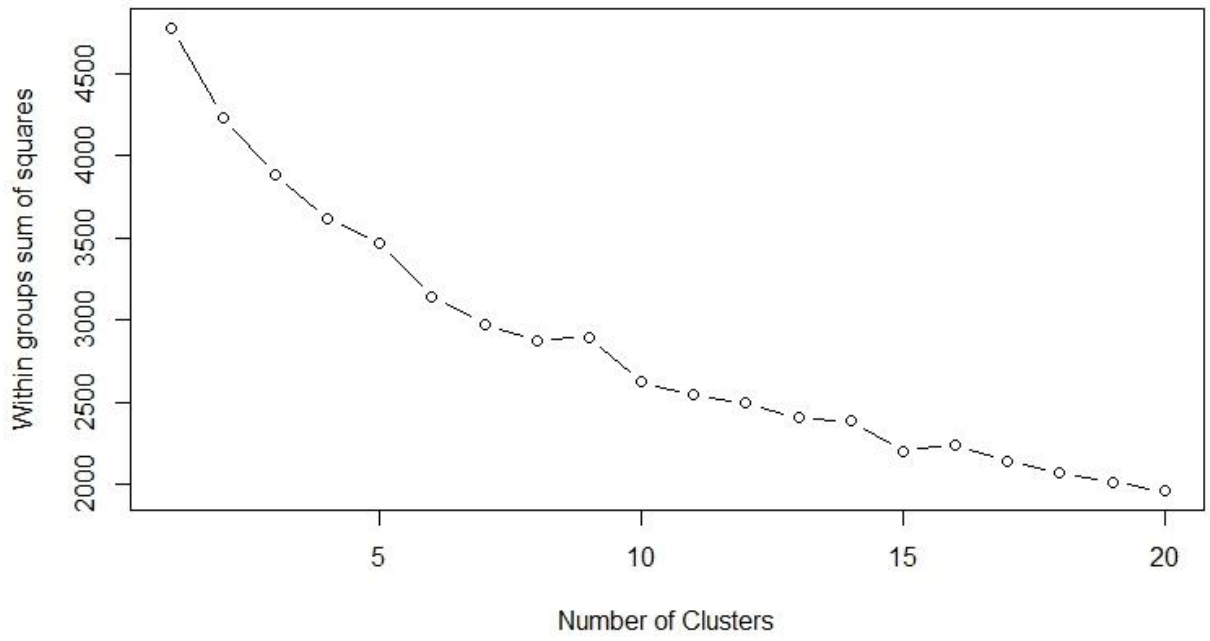

Figure IV-5 Choix du nombre de cluster

Nous pouvons conclure que si nous conservons un nombre de clusters = 5, nous devrions pouvoir obtenir de bons clusters avec une bonne homogénéité en eux-mêmes. Il existe la fonction NbClust() du package du même nom, qui offre à l'utilisateur le meilleur schéma de regroupement parmi les différents résultats.

- La première étape Map initialise l'algorithme en répartissant aléatoirement les observations en k classes puis gère l'affectation des observations aux centres les plus proches, **[36.37.38.39]**
- L'étape Reduce calcule les nouveaux centres

#Effectuons un test sur données simulées Simulation de 5 centres

> set.seed(1)

P=do.call(rbind,rep(list(matrix(rnorm(10, sd = 10),ncol=2)),20))+

matrix(rnorm(200), ncol =2) out = kmeans.mr(to.dfs(P), num.clusters = 5, num.iter = 8)

> plot(P)

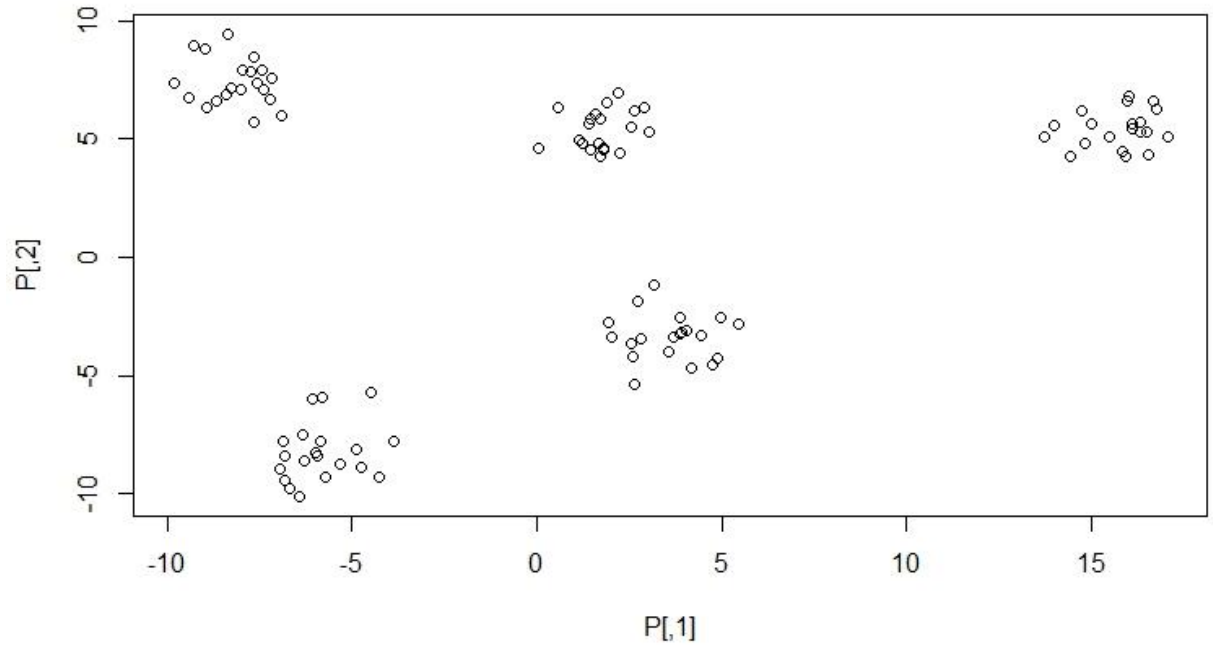

Figure IV-6: Simulation de 5 clusters

IV-8 Interprétation des résultats

Plusieurs techniques sont proposées pour le problème général de la classification. Ils différent par les mesures de proximités qu'ils utilisent ,la nature des données qu'ils traitent et l'objectifs finale de la classification, chacune de ces technique possède ses points forts et ses points faibles, les méthodes hiérarchiques ascendantes sont utilisées en cas des données de petite taille car la complexité est très élevée , et Si des problèmes de temps d'exécution se posent Lourdeur des calculs dès qu'on a un nombre de données important , alors c'est les méthodes des K-means qui sont utilisées L'avantage de ces algorithmes est avant tout leur grande simplicité.

- Tend à réduire l'erreur quadratique.
- Applicable à des données de grandes tailles

IV-9 Conclusions

Après avoir analysé les résultats des tests des algorithmes de clustering et les avoir exécutés sous différents facteurs et situations, les conclusions suivantes sont obtenues :

- À mesure que le nombre de clusters, k, augmente, les performances de l'algorithme K-means et meilleures que celles de l'algorithme de clustering hiérarchique,
- A mesure que la valeur de k augmente, la précision de la classification hiérarchique devient meilleure,

- L’algorithme K-means ont moins de qualité (précision) que la classification hiérarchique,
- Les deux algorithmes présentent une certaine ambiguïté dans certaines données (bruyantes) lorsqu'elles sont regroupées,
- La qualité de l’ algorithmes K-means devient très bonne lors de l'utilisation d'un vaste ensemble de données.
- Le clustering hiérarchique donne de bons résultat lors de l'utilisation de petits ensembles de données
- Les algorithmes de partitionnement K-means ont recommandés pour un vaste ensemble de données, tandis que les algorithmes de clustering hiérarchique sont recommandés pour un petit ensemble de données,
- Le clustering hiérarchique donne de meilleurs résultats par rapport aux algorithmes K-means lors de l'utilisation d'un ensemble de données aléatoires et vice versa,
- Les algorithmes K-means sont très sensibles au bruit dans l'ensemble de données. Ce bruit empêche l'algorithme de regrouper un objet dans son cluster approprié. Cela affectera les résultats de l'algorithme.

Conclusion Générale

Dans ce mémoire, on a essayé de trouver une approche qui pourrait simplifier le traitement des données de grand masse et sur monter aussi la charge des calculs surtout quand on a des moyens limités. L'idée de cette approche consiste à appliquez quelque technique de fouille de données sur le big data (données massive) par la méthode de k-means et hiérarchies (CAH). les méthodes hiérarchiques ascendantes sont utilisées en cas des données de petite taille car la complexité est très élevée, et Si des problèmes de temps d'exécution se posent Lourdeur des calculs dès qu'on a un nombre de données important, alors c'est les méthodes des K-means qui sont utilisées L'avantage de ces algorithmes est avant tout leur grande simplicité.

· Tend à réduire l'erreur quadratique.
· Applicable à des données de grandes tailles

Dans le cadre de travaux futurs, des comparaisons entre ces deux algorithmes (ou peut-être d'autres algorithmes) pourront être tentées en fonction de facteurs différents autres que ceux considérés dans cet mémoire. Un facteur important est la normalisation. La comparaison entre les résultats d'algorithmes utilisant des données normalisées ou des données non normalisées donnera des résultats différents.

Bien entendu, la normalisation affectera les performances de l'algorithme et la qualité des résultats. Une autre approche peut envisager d'utiliser des algorithmes de regroupement de données dans des applications telles que la reconnaissance d'objets et de caractères ou la récupération d'informations qui concerne le stockage automatique et la récupération de documents.

Bibliographie

[1] G. Piatetsky-shapiro, Data Mining And Knowledge Discovery 1996 to 2005: Overcoming The Hype And Moving From «University» To «Business» And «Analytics», Data Mining And Knowledge Discovery, 15(1), 99-105.

[2] The Gartner Group, www.gartner.com.

[3] D. Hand, h. Mannila et P. Smyth, Principles Of Data Mining,Mit Press, cambridge, ma, 2001.

[4] P. Cabena, P. Hadjinian, r. Stadler, j. Verhees et a. Zanasi,

Discovering data mining: From Concept To Implementation, Prentice Hall, Upper Saddle River, nj, 1998.

[5] E-G. Talbi, Fouille De Données (Data Mining) : Un tour d'horizon, Laboratoire D'informatique Fondamentale De Lille.

[6] Abdelmalek Amine, laboratoire gecode - Universté De Saida, Cours Data Mining, Ecole D'hiver Sur Les Applications De L'informatique Industrielle, Réseaux Et Génie Logiciel, Université D'oran, 09-12 Décembre 2013

[**6**] Oded maimon, Lior rokach, Data Mining and Knowledge Discovery handbook (second Edition), springer, isbn 978-0-387-09822-7, e-isbn 978-0-387-09823-4, 2010.

[7] S. Prabhu, n. Venkatesan, Data Mining and Warehousing, New age international (p) ltd., publishers, new delhi, 2007

[8] G. Calas, Etudes Des Principaux Algorithmes De Data Mining, Specialisation Sciences Cognitives Et Informatique Avancee, France.

[9] R.Gilleron, M. Tommasi, Découverte De Connaissances A Partir De Données, 2000.

[10] K.Teknomo, What Is Nearest Neighbors Alghorithm?

http://people.revoledu.com/kardi/tutorial/knn/what-is-k-nearestneighbor- algorithm.html, 2006.

[**11**] C. Scharff, Méthode Des K Plus Proches Voisins, ifi, 2004.

[12] L'encyclopedie En Ligne Wikepedia 2009.

[13] Rapporté de http://interstices.info/encart.jsp?id=c_41867&encart=3&size=600,500.

[14] La Détection Automatique De Clusters, Rapporté de O. El ganaoui, m. Perrot, Segmentation Par Régions: Une Méthode Qui Utilise La Classification Par Nuées Dynamiques Et Le Principe D'hysteresis, 31 Décembre 2004.

[15] Http://Www.Eisti.Fr/~Lassi/Pfe/Id/Datamining/Site/Detection.Htm

[16] J. Han, M. Kamber, and J. Pei. Data mining : Concepts And Techniques. Morgan Kaufmann Pub, 2011.

[17] M. Kantardzic. Data mining : Concepts, Models, Methods, And Algorithms. Wileyinterscience, 2003.

[18] P. Preux. Fouille de données, Notes De Cours. Disponible sur internet, 2006.

[19] Hurwitz, J, Nugent, A, Halper, F, and Kaufman, M (2013). Big Data For Dummies? Ebook.https://eecs.wsu.edu/~yinghui/mat/courses/fall%202015/resources/Big%20data% 20for%20dummies.pdf.

[20]. Popescu, A, And Bacalu, Am (2012). Big Data Causes Concern And Big Confusion. A Big Data Definition To Help Clarify The Confusion. Rapport De Thor Olavsrud (Cio) Pourit World Resultants D'une Enquête Sur Big Data.

[21] Fermigier, S (2012). Big Data and Open Source: Une Convergence Inevitable. Livre Blanc (http://Fermigier.Com/Blog/2012/03/New-Whitepaper-Big-Data-Open-Source/)

[22] Guillaume Cleuziou, Une Méthode De Classification Non-Supervisee Pour L'apprentissage De Regles Et La Recherche D'information, Décembre 2004, Université D'orleans Page 07, Pages11-15

[23] Guénaël CABANES, Classification Non Supervisée A Deux Niveaux Guidée Par Le Voisinage Et La Densite, 03/12/10, Universite Paris 13 Pages 1

[24] Nicoleta Rogovschi, Classification A Base De Modeles De Melanges Topologiques Des Données Categorielles Et Continues, Université Paris 13 - Institut Galilee Laboratoire D'informatique De Paris Nord Umr 7030 Du Cnrs

[25] Bounneche Meriem Dorsaf, Reduction De Donnees Pour Le Traitement D'images, 2009, Université Mentouri Constantine, Pages 04 – 07

[26] Losee, R.M. (1998). Text Retrieval And Filtering Analytic Models Of Performance. kluwer academic publishers.

[27] A. K. Jain, M. N. Murty, and P. J. Flynn. Data Clustering: A Review. Acm Computing surveys, 31(3):264_323, 1999.

[28] Spath H. Clustering Algorithms. John Wiley and Sons, New York, NY, 1975.

[29] J. A. Hartigan and M. A. Wong. Algorithm AS 136 : A K-Means Clustering algorithm. Applied Statistics, 28(1) :100_108, 1979.

[30] E. Forgy. Cluster Analysis Of Multivariate Data : Eficiency Versus Interpretability Of classifications. Biometrics, 21 :768_780, 1965.

[31] Inderjit S. Dhillon, Subramanyam Mallela, and Rahul Kumar. Enhanced Word Clustering For Hierarchical Text Classi_cation. In kdd '02: Proceedings Of The Eighth Acm Sigkdd International Conference On Knowledge Discovery And Data Mining, pages 191_200, new york, ny, usa, 2002. Acm.

[32] Clefs CEA N° 64 - Juin 2017 Voyage Au Coeur Du Big Data.

[33] Https://Fr.Wikipedia.Org/Wiki/Rstudio

[34] Https ://Www.Fun-Mooc.Fr/C'est Une Enquete De Satisfactions Dans Un Hôpital, Récupère Lors D'un Cours De Fun (France Unité Numérique).(Il s'agit d'une études évaluant la qualité de relation et la quantité d'information reçue par le patient lors de son séjour a l'hôpital).

[35] Https ://Cran.R-Project.Org/Web/Packages/Nbclust/Index.Html

[36] Https://Www.Math.Univ-Toulouse.Fr/~Besse/Wikistat/ Pdf/St-Tutor5-R-Mapreduce.Pdf

[37] Http://Eric.Univ-Lyon2.Fr/~Ricco/Cours

[38] Http://Chirouble.Univ-Lyon2.Fr/~Ricco/Cours/Index.Html

[39] http://data.mining.free.fr.

[40] Https://Www.Researchgate.Net/Publication/293061584

[**41**] Abbas ,O.A., Jordan, "Comparisons Between Data Clustering Algorithms, "The International Arab Journal of Information Technology, vol. 5, no. 3, pp. 320-326, Jul. 2008.

[**42**] Dr. Manju Kaushik, Mrs. Bhawana Mathur, "Comparative Study of K-Means and Hierarchical Clustering Techniques"

Résumé

Fouille de données et données massives sont deux concepts différents. Les données massives font référence à une grande quantité de données, les techniques de fouille de données ont montrés de bonnes performances dans la détection d'intrusion afin de trouver le bon technique qui nous permet d'obtenir les meilleur performances , nous avons appliquez et comparé deux technique de fouille de données sur le big data classification k-means et le classification hiérarchique après avoir analysé les performance des différent classification , conclu que le classification k-means est le meilleur technique par rapport aux autres technique malgré cette analyse on ne peut pas le définir comme étant le meilleur technique en raison de l'existant d'autre technique de sélection d'attribut qui n'ont pas été utilise dans notre étude comparative.

Mots clés : Données massives, Fouille de données, k-means, classification hiérarchique

Abstract

Data Mining and Big Data are two different concepts. Big Data refers to a large amount of data, data mining techniques have shown good performance in intrusion detection in order to find the right technique that allows us to obtain the best performance, we compared two techniques of data mining on big data clustering k-means and hierarchical clustering after analyzing the performance of the different classifications we concluded that the clustering k-means is the best technique compared to other techniques despite this analysis we cannot define it as the best technique due to the existence of other attribute selection techniques that were not used in our comparative study.

Keywords: Big Data, Data Mining, k-means, hierarchical clustering

ملخص

التنقيب في البيانات والبيانات الضخمة مفهومان مختلفان. تشير البيانات الضخمة إلى كمية كبيرة من البيانات ، وقد أظهرت تقنيات استخراج البيانات أداءً جيدًا في اكتشاف التسلل من أجل العثور على التقنية الصحيحة التي تتيح لنا الحصول على أفضل أداء ، وقمنا بتطبيق مقارنة تقنيتين من التنقيب عن البيانات على البيانات الضخمة باستعمال التجميع المراكز المتنقلة و التجميع الهرمي بعد تحليل أداء التصنيفات المختلفة خلصنا إلى أن التجميع باستعمال التجميع المراكز المتنقلة هو أفضل تقنية مقارنة بالتقنيات الأخرى على الرغم من هذا التحليل لا يمكننا تعريفه على أنه أفضل تقنية بسبب وجود تقنيات اختيار السمات الأخرى التي لم يتم استخدامها في دراسة هذه المقارنة.

كلمات مفتاحية: البيانات الضخمة، التنقيب عن البيانات، طريقة k-means، التجميع الهرمي