Université Abou Bekr Belkaid
Tlemcen Algérie
جامعة أبي بكر بلقايد
تلمسان الجزائر

**République Algérienne Démocratique et Populaire**
**Université Abou Bakr Belkaid– Tlemcen**
**Faculté des Sciences**
**Département d'Informatique**

**Mémoire de fin d'études**

**Pour l'obtention du diplôme de Licence en Informatique**

# Thème

# Clustering Hiérarchique de données à base de Ward

**Réalisé par :**

- **Tabet aoul Walid Houcine**
- **Ziani Soheyb**

*Présenté le 27Juin 2013 devant la commission d'examination composée de MM.*

- *Mr.Hadjila Fethallah* *(Examinateur)*
- *Mr.Merzoug Mohamed* *(Examinateur)*
- *Mr.Belabed Amine* *(Examinateur)*

Année universitaire : 2012-2013

# Table de matières

# Remerciement

Avec l’aide du Dieu le tout puissant nous avons pu achever ce travail.

Nous adressons nos sincères et chaleureux remerciements aux personnes qui ont contribué à l’élaboration de ce mémoire et nous ont aidé à le réaliser même par des conseils.

En premier lieu, nous tienderenos à remercier Mr. HADJILA.Fethallah, professeur à l'université de Chetouane Tlemcen. En tant que Directeur de mémoire, il nous a guidé dans notre travail et nous a aidé à trouver des solutions pour avancer et s'est toujours montré à l'écoute et très disponible tout au long de la réalisation de ce mémoire, ainsi pour l'inspiration, l'aide et le temps qu'il a bien voulu nous consacrer et sans lui ce mémoire n'aurait jamais vu le jour.

Enfin, nous adressons nos plus sincères remerciements à tous nos proches et amis, qui nous ont toujours soutenue et encouragée au cours de la réalisation de ce mémoire. Merci à tous et à toutes.

# Table de figure

# Introduction générale

## Contexte et problématique

Le regroupement se base sur l'idée de déterminer la classification dans un ensemble de données non étiquetées. Mais comment décider ce qui constitue un regroupement de bon? On peut montrer qu'il n'y a pas d'absolu "meilleur" critère qui serait indépendant de l'objectif final de la classification. Par conséquent, c'est l'utilisateur qui doit fournir ce critère, de telle sorte que le résultat du regroupement répondra à leurs besoins.

L'apprentissage automatique est employé pour étiqueter correctement des exemples. L'apprenant doit alors trouver ou approximer la fonction qui permet d'affecter la bonne étiquette à ces exemples. l'objectif de définir des groupes d'exemples tels que la distance entre exemples d'un même groupe soit minimale et que la distance entre groupes soit maximale (ces deux contraintes vont dans des sens opposés et c'est le meilleur compromis qui doit être trouvé) est de trouver des regroupements de manière naturelle ;et bien on parle de *l'apprentissage supervisé* quand les classes sont bien connues et même les exemples sont fournis avec l'étiquette de leur classe par contre *l'apprentissage non supervisé* est tout à fait le contraire ou on dispose des exemples sans étiquettes et d'une distance définie sur le langage de description des exemples et l'espace de description est un espace vectoriel numérique (en pratique,$R^n$) dans lequel chaque dimension correspond à un attribut distinct. L'exemple est décrit par un vecteur d'attributs à valeurs réelles, le troisième type d'apprentissage *semi-supervisé* est une classe de techniques utilisé a un ensemble de données étiquetées et non-étiquetés.

L'apprentissage non supervisé ou « clusternig » constitue le problème principal de ce travail, en effet nous distinguons plusieurs approches dans ce domaine, la classification hiérarchique est considérée comme l'une des approches les plus répandues.

**Contribution**

L'objectif de ce mémoire est d'étudier et d'implémenter une approche qui fait partie de la classification hiérarchique. En particulier nous implémentons la classification ascendante base distance de Ward.

# Plan du mémoire

Le reste de mémoire est structuré comme suit :

**Chapitre 1** : Il permet de présenter les différentes classes d'apprentissage automatique, en particulier nous introduisons l'apprentissage supervisé, non supervisé (clustering), et semi supervisé. En outre nous montrons un ensemble de techniques de clustering telles que K-means, le clustering hiérarchique ascendant, E.M…

**Chapitre 2** : Il permet de montrer la conception et implémentation de notre prototype.

## I.1 Introduction

Notre monde est plein du défis et des informations donc l'être humain est toujours sensé d'apprendre et de rechercher pour qu'il puisse attribuer à chaque chose sa définition et sa place alors…L'apprentissage est l'acquisition de savoir-faire, c'est-à-dire le processus d'acquisition de pratiques, de connaissances, compétences, d'attitudes ou de valeurs culturelles, par l'observation, l'imitation, l'essai, la répétition, la présentation. En générale il y'a deux types d'apprentissage, l'apprentissage ″ par cœur ″ qui consiste à mémoriser une information sans tenir compte de son sens. Autrement dit, mémoriser uniquement la forme. Cela signifie que l'information sera stockée sans être associée à son sens, et l'apprentissage par généralisation c'est pouvoir extraire des généralités à partir des exemples dont on dispose à partir d'une attribution d'une classe spécifique à un objet donné.

La difficulté dans les systèmes informatiques réside dans la généralisation d'un certain nombre de données comme (textes, images, vidéos, musiques ...), ou il est par contre facile de les mémoriser.

Donc, L'apprentissage automatique est la tentative pour connaitre ces défis et bien modéliser des connaissances liant (texte et image etc..). Donc dans ce chapitre on représente les techniques du clustering.

## I.2 Définitions d'apprentissage automatique :

L'apprentissage automatique fait référence au développement, à l'analyse et à l'implémentation de méthodes qui permettent à une machine d'évoluer grâce à un processus d'apprentissage, et ainsi de remplir des tâches qu'il est difficile ou impossible de remplir par des moyens algorithmiques plus classiques.

Des systèmes complexes peuvent être analysés, y compris pour des données associées à des valeurs symboliques (ex: sur un attribut numérique, non pas simplement une valeur numérique, *juste un nombre*, mais une valeur probabilisée, c'est-à-dire un nombre assorti d'une probabilité ou associé à un intervalle de confiance) ou un ensemble de modalités possibles sur un attribut numérique ou catégoriel. L'analyse peut même concerner des données présentées sous forme de

graphes ou d'arbres, ou encore de courbes (par exemple, la courbe d'évolution temporelle d'une mesure ; on parle alors de *données continues*, par opposition aux *données discrètes* associées à des attributs-valeurs classiques) [1] .

En tous les cas, il consiste à utiliser des ordinateurs pour optimiser un modèle de traitement de l'information selon certains critères de performance à partir d'observations, que ce soit des données-exemples ou des expériences passées .Lorsque l'on connait le bon modèle de traitement à utiliser, alors pas besoin de faire de l'apprentissage.

L'apprentissage automatique peut être utile lorsque :

● On n'a pas d'expertise sur le problème .Premier exemple : ‟robot navigant sur Mars”.

● On a une expertise, mais on ne sait pas comment l'expliquer .Premier exemple : ‟ reconnaissance de visages”.

● Les solutions au problème changent dans le temps .Premier exemple : ‟ routage de Paquets”.

● Les solutions doivent être personnalisées .Premier exemple : ‟biométrie”.

● Il y'a deux phases d'apprentissage

1) On présente des exemples au système.
2) Le system « apprend » à partir des exemples.

Donc le système modifie graduellement ses paramètres ajustables pour que sa sortie ressemble à la sortie désirée.

L'apprentissage automatique (*machine learning* en anglais), un des champs d'étude de l'intelligence artificielle, est la discipline scientifique concernée par le développement, l'analyse et l'implémentation de méthodes automatisables qui permettent à une machine (au sens large) d'évoluer grâce à un processus d'apprentissage, et ainsi de remplir des tâches qu'il est difficile ou impossible de remplir par des moyens algorithmiques plus classiques.[2]

## I.3 Applications

L'apprentissage automatique est utilisé pour doter des ordinateurs ou des machines de systèmes de : perception de leur environnement : vision, reconnaissance d'objets (visages, schémas, langages naturels, écriture, formes syntaxiques, etc.) ; moteurs de recherche ; aide aux diagnostics, médical notamment, bio-informatique, schéma informatique ; interfaces cerveau-machine ; détection de fraudes à la carte de crédit, analyse financière, dont analyse du marché boursier ; classification des séquences d'ADN ; jeu ; génie logiciel ; sites Web adaptatifs ou mieux adaptés ; locomotion de robots ; etc.

Exemples :

- ✓ Un système d'apprentissage automatique peut permettre à un robot ayant la capacité de bouger ses membres mais ne sachant initialement rien de la coordination des mouvements permettant la marche, d'apprendre à marcher. Le robot commencera par effectuer des mouvements aléatoires, puis, en sélectionnant et privilégiant les mouvements lui permettant d'avancer, mettra peu à peu en place une marche de plus en plus efficace.
- ✓ La reconnaissance de caractères manuscrits est une tâche complexe car deux caractères similaires ne sont jamais exactement égaux. On peut concevoir un système d'apprentissage automatique qui apprend à reconnaître des caractères en observant des « exemples », c'est-à-dire des caractères connus.

## I.4 Types d'apprentissage

Les algorithmes d'apprentissage peuvent se catégoriser selon le type d'apprentissage qu'ils emploient :

**●L'apprentissage supervisé**

**●L'apprentissage non-supervisé**

**●L'apprentissage par renforcement (semi –supervisé)**

- Il est Illustré selon le schéma suivant :

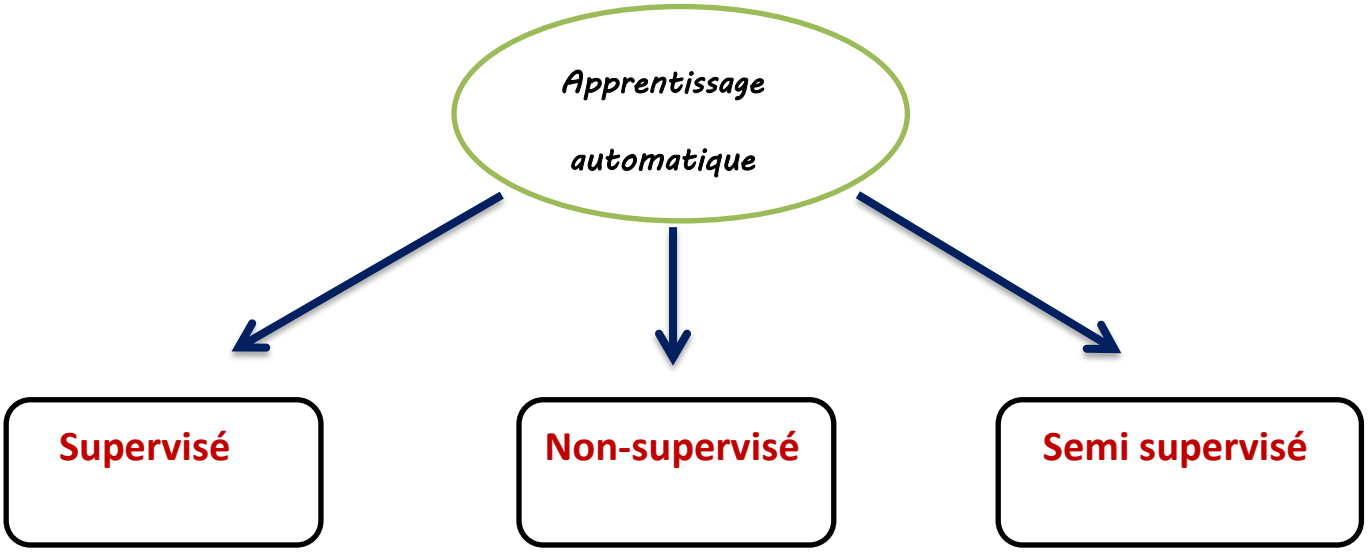



**Figure I .1 :** Différents types d'apprentissage

Une donnée (un exemple) est un couple (x, f(x)) ou x est la valeur d'entrée et f(x) la valeur de sortie (inconnue si apprentissage non-supervisé).

**Le supervisé :**

■ Les données sont composées d'exemples et de contre-exemples (entrées et sortie de la fonction / appartenance ou non à la classe).

■ Donner la classe de la donnée.

**Le non supervisé :**

Qui recherche de similarités dans les données : on ne connaît pas les classes a priori

**Le Semi-supervisé :**

C’est Le troisième type qui couvre l’espace existant ente le supervisé et le non supervisé.

## I.5 L’apprentissage supervisé

### I.5.1 Introduction

La notion de prédiction fait référence à une stratégie particulière d'Apprentissage Automatique (AA), appelée apprentissage supervisé. Ce domaine d'étude à part entière, peut être considéré comme une sous-thématique de l'Intelligence Artificielle (IA). De façon synthétique, l'apprentissage supervisé consiste à faire émerger d'un ensemble de données d’entrainement pré-classifiées, les caractéristiques nécessaires et suffisantes pour permettre de classer correctement une nouvelle donnée. Dans ce type d'approche, les classes sont connues à l'avance, cette connaissance est utilisée dans le processus d'apprentissage. [6]

### I.5.2 définition

Si les classes sont prédéterminées et les exemples connus, le système apprend à classer selon un modèle de classement ; on parle alors d'apprentissage supervisé (ou d'analyse discriminante). Un expert (ou oracle) doit préalablement étiqueter des exemples. Le processus se passe en deux phases. Lors de la première phase (hors ligne, dite d'apprentissage), il s'agit de déterminer un modèle des données étiquetées. La seconde phase (en ligne, dite de test) consiste à prédire l'étiquette d'une nouvelle donnée, connaissant le modèle préalablement appris. Parfois il est préférable d'associer une donnée non pas à une classe unique, mais une probabilité d'appartenance à chacune des classes prédéterminées (on parle alors d'apprentissage supervisé probabiliste).

**Exemple** : L'analyse discriminante linéaire ou les SVM en sont des exemples typiques. Autre exemple : en fonction de points communs détectés avec les symptômes d'autres patients connus (les « exemples »), le système peut catégoriser de nouveaux patients au vu de leurs analyses médicales en risque estimé (probabilité) de développer telle ou telle maladie.[3]

### I.5.3 Principe

1- Un expert est employé pour étiqueter correctement des exemples.

2- L'apprenant doit alors trouver ou approximer la fonction qui permet d'affecter la bonne étiquette à ces exemples.

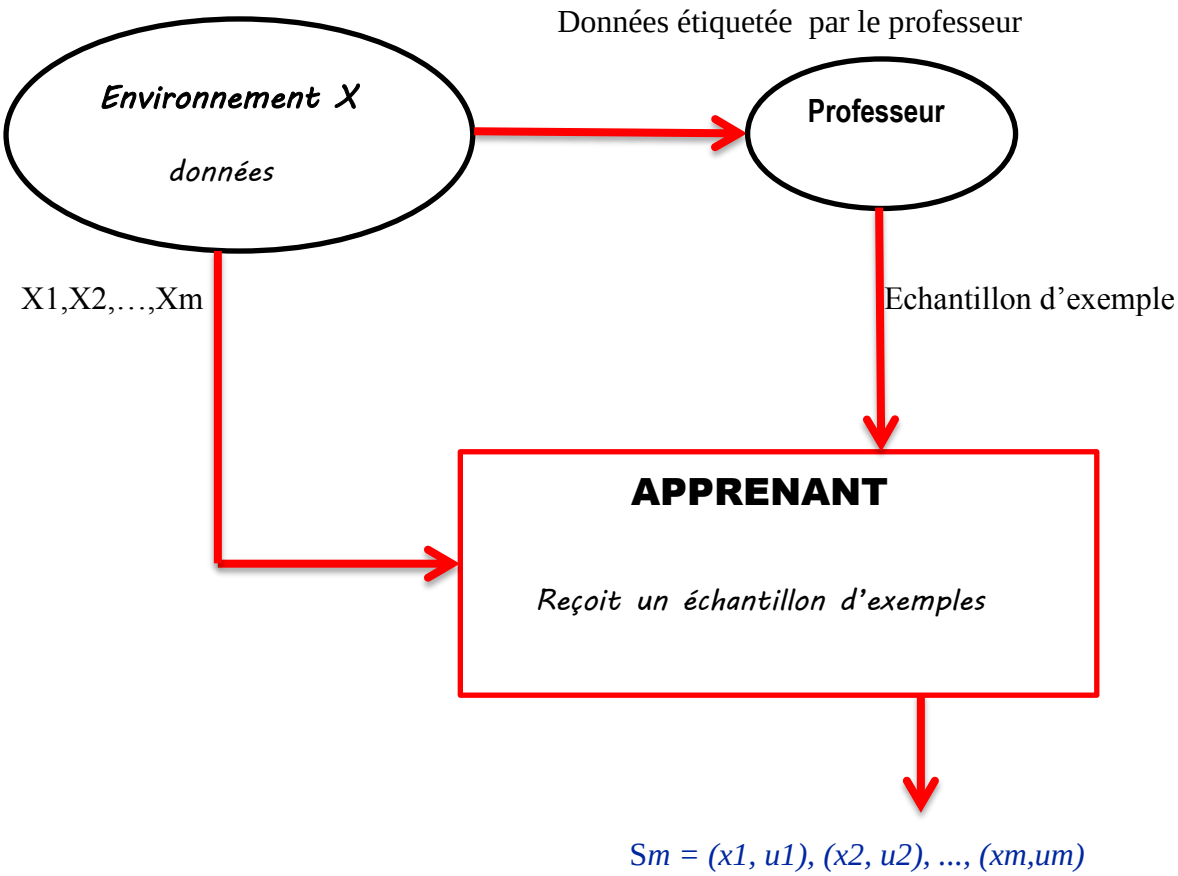



**Figure I.2 :** Système de fonctionnement de l'apprentissage supervisé

### I.5.4 Les Différents buts d'apprentissage supervisé

**I.5.4.1** Buts principaux

- Approximer au mieux la sortie désirée pour chaque entrée observée.
- Obtenir un « bon » modèle : la prévision obtenue est proche de la vraie valeur.
- Obtenir rapidement un modèle rapide : temps de construction du modèle et temps nécessaire à l'obtention d'une prévision.

- Pouvoir garantir les performances : avec une probabilité de 1 – r ; la prévision sera bonne à peu près.

**I.5.4.2** Buts annexes **:**

- obtenir un modèle compréhensible : comment le modèle prend il la décision ?
- obtenir un modèle modifiable : pouvoir prendre en compte de nouvelles données ; s'adapter à un environnement changeant, etc.

**I.5.4.3** Buts mathématiques **:**

- apprendre une projection entre des observations X en entrée et des valeurs associées Y en sortie. [8]

### I.5.5 Quelques algorithmes d'apprentissage supervisé

La plupart des algorithmes d'apprentissage supervisés tentent de trouver un **modèle** (une fonction mathématique) qui explique le lien entre des données d'entrée et les classes de sortie. Ces jeux d'exemples sont donc utilisés par l'algorithme.

Il existe de nombreuses méthodes d'apprentissage supervisé parmi eux :

- Méthode des k plus proches voisins.
- Machine à vecteurs de support .
- Réseau de neurones.
- Arbre de décision.
- Classification naïve bayésienne.
- Inférence grammaticale.

## I.6 Apprentissage non supervisé (Clustering)

### I.6.1 Introduction

Quand le système ou l'opérateur ne disposent que d'exemples, mais non d'étiquettes, et que le nombre de classes et leur nature n'ont pas été prédéterminés, on parle d'apprentissage non supervisé ou clustering. Aucun expert n'est requis. L'algorithme doit découvrir par lui-même la structure plus ou moins cachée des données. Le partitionnement de données, data clustering en anglais, est un algorithme d'apprentissage non supervisé.

Le système dans l'espace de description (la somme des données) doit cibler les données selon leurs attributs disponibles, pour les classer en groupe homogènes d'exemples. La similarité est généralement calculée selon une fonction de distance entre paires d'exemples. C'est ensuite à l'opérateur d'associer ou déduire du sens pour chaque groupe et pour les motifs (patterns en anglais) d'apparition de groupes, ou de groupes de groupes, dans leur « espace ». Divers outils mathématiques et logiciels peuvent l'aider. On parle aussi d'analyse des données en régression (ajustement d'un modèle par une procédure de type moindres carrés ou autre optimisation d'une fonction de coût). Si l'approche est probabiliste (c'est-à-dire que chaque exemple, au lieu d'être classé dans une seule classe, est caractérisé par un jeu de probabilités d'appartenance à chacune des classes), on parle alors de « soft clustering » (par opposition au « hard clustering »). Cette méthode est souvent source de sérénité. [5]

### I.6.2 Définition

La méthode non-supervisée appelé aussi « clustering » ne nécessite pas d'avoir identifié au préalable des classes correspondant aux différents clusters. L'objectif du clustering est de former des clusters tels que chaque cluster soit homogène (les éléments de chaque cluster doivent être le plus similaires possible) et les clusters doivent être hétérogènes entre eux (deux clusters doivent être le plus dissimilaires possible). La plupart des méthodes utilisent une mesure pour calculer la qualité d'un clustering en se basant sur une distance intra-cluster et une distance inter-clusters. Les algorithmes cherchent alors à obtenir un clustering de qualité optimale, correspondant à une distance intra-cluster faible et une distance inter-clusters élevée. Traditionnellement, les techniques de clustering sont divisées en trois familles: les méthodes de

partitionnement, les approches hiérarchiques et la classification conceptuelle. Il existe aussi des méthodes récentes qui s'appuient sur la notion de motif (en anglais, item set) afin de ne pas définir une distance entre les objets.

Exemple : Pour un épidémiologiste qui voudrait dans un ensemble assez large de victimes de cancers du foie tenter de faire émerger des hypothèses explicatives, l'ordinateur pourrait différencier différents groupes, que l'épidémiologiste chercherait ensuite à associer à divers facteurs explicatifs, origines géographique, génétique, habitudes ou pratiques de consommation, expositions à divers agents potentiellement ou effectivement toxiques (métaux lourds, toxines telle que l'aflatoxine, etc.).[4]

Il en découle que la mise en œuvre de la plupart des techniques de classification ne nécessite que des notions mathématiques relativement élémentaires parlons sur la distance euclidienne usuelle :

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

### I.6.3 quelques méthodes de la classification non supervisé

Il y a plusieurs et on distingue 3 sortes de méthodes principales généralement connus :

#### I.6.3.1 Méthodes de partitionnement

Les approches de classification par partitionnement permettent de subdiviser l'ensemble des individus en un certain nombre de classes en employant une stratégie d'optimisation itérative dont le principe général est de générer une partition initiale, puis de chercher à l'améliorer en réattribuant les données d'une classe à l'autre. Il n'est bien entendu pas souhaitable d'énumérer toutes les partitions possibles. Ces algorithmes recherchent donc des maxima locaux en optimisant une fonction objectif traduisant le fait que les individus doivent être similaires au sein d'une même classe, et dissimilaires d'une classe à une autre. Les classes de partition final, prises deux à deux, sont d'intersection vide est représentée par noyau.

❖ Principe :

Cherche la meilleure partition en K classes disjointes des données, le nombre de classes (clusters ou groupes) K étant fixé a priori. Les approches par partitionnement utilisent un processus itératif fonction de nombre K qui consiste à affecter chaque individu à la classe la plus proche au sens d'une distance –ou d'un indice de similarité-en optimisant une certaine fonction objectif.

❖ Différentes approches :

● **Centres mobiles** : La méthode des centres mobiles due à Forgy [Forgy, 1965] est la plus classique et celle qui reste très utilisée. Elle procède comme suit : dans une première étape, elle consiste à tirer aléatoirement k individus de la population. Ces individus représentent les centres provisoires des k classes qui formeront la partition initiale. Ensuite, les autres individus sont regroupés autour de ces k centres en affectant chacun d'eux au centre le plus proche. L'étape suivante consiste à recalculer les k nouveaux centres (dites aussi centroïdes ou centres de gravité) des k classes, sachant qu'un centre n'est pas nécessairement un individu de la population. Le processus est répété plusieurs fois jusqu'à stabilité des centres des classes (les centres ne bougent plus).

● **K –means(Mc Queen) :** La procédure est un moyen simple et facile de classer un ensemble à travers un certain nombre de clusters (grappes supposer k) données fournies fixé a priori. L'idée principale est de définir k centres de gravité, un pour chaque cluster. Ces centres de gravité qui devraient être placés d'une manière rusée en raison de son emplacement différent provoquent résultat différent. Donc, le meilleur choix est de les placer autant que possible loin les uns des autres. L'étape suivante consiste à prendre chaque point appartenant à un ensemble de données et l'associer au centre de gravité le plus proche. Lorsqu'aucun point n'est en cours, la première étape est terminée et un groupage début est fait. A ce stade, nous avons besoin de recalculer k nouveaux centres de gravité comme barycentres des groupes issus de l'étape précédente. Après nous avons ces nouveaux centres de gravité k, une nouvelle liaison doit être fait entre les mêmes données des points de consigne et le nouveau centre de gravité le plus proche

● **Nuées dynamiques (Duday) :** La méthode des nuées dynamiques largement développé par Diday dans [Diday, 1971] se distingue principalement des approches précédentes par le mode de représentation des classes appelé aussi noyau. Ce dernier peut être son centre de gravité (dans ce cas nous retrouvons l'approche des centres mobiles), un ensemble d'individus (l'approche des k-médianes avec un seul individu), une distance (l'approche des distances adaptatives [Diday and Govaert, 1977]), une loi de probabilité (la décomposition de mélanges [Schroeder, 1976]), etc.

● **K-représentants (k-médoids) :** Les méthodes k-médianes présentent l'avantage d'être applicable à tout type de données et sont dans l'ensemble plus robustes aux points aberrants que les méthodes des k-moyennes, d'autant qu'elles recourent aux médianes (medoïdes) plutôt qu'aux moyennes (centroïdes) pour évaluer la distance aux centres. On peut même parler sur k-modes, k-prototypes.

● **Réseaux de kohonen :** proposé en 1997 est un procédé d'auto-organisation qui cherche à projeter des données multidimensionnelles sur un espace de faible dimension (en général de dimension 2), appelé carte. Cette réduction de dimension (ou aussi projection) permet d'obtenir un partitionnement des individus en groupes similaires, tout en préservant au mieux la structure topologique des données.

● **Méthodes basés sur la notion de densité** : L'idée des méthodes par densité est de définir une classe comme étant un ensemble d'individus de forme quelconque, mais "dense" selon un critère de voisinage et de connectivité. Ces méthodes sont fondées sur les concepts de densité, noyau, point limite, accessibilité et connectivité.

### I.6.3.2 Méthodes Mixtes

Cette approches combinent efficacement les avantages des deux autres types de méthodes et permettent d'en annuler les inconvénients ; Ces méthodes, qui combinent les méthodes de collecte et d'analyse de-données qualitatives et quantitatives, sont de plus en plus populaires et, au fil des ans, plusieurs termes ont été proposés pour nommer ces combinaisons; études triangulées ou multi méthodes, ou encore méthodes mélangées, intégrées, multiples ou mixtes. Johnson ·(2007) a effectué une revue critique de la littérature scientifique sur les méthodes

mixtes et a proposé une définition de ces méthodes pour la recherche. Cette revue de littérature suggère de définir les méthodes mixtes pour l'évaluation de programme de la manière suivante. Une évaluation mixte est un type d'évaluation dans lequel un expert ou une équipe d'experts combine les méthodes qualitatives et quantitatives d'évaluation (approches et (ou) devis et (ou) techniques de collecte et d'analyse de données) dans le but d'approfondir la compréhension et la corroboration des résultats d'évaluation des programmes.

### I.6.3.3 Méthodes hiérarchiques

**Définition**

Ici se produisent une séquence de partitions emboitées d'hétérogénéités croissantes de la plus fine à la plus grossière, conduisent à des résultats sous forme d'arbre hiérarchique indicé connu om de aussi sous le nom de dendrogramme, qui visualise ce système de classes organisées par inclusion. [12]

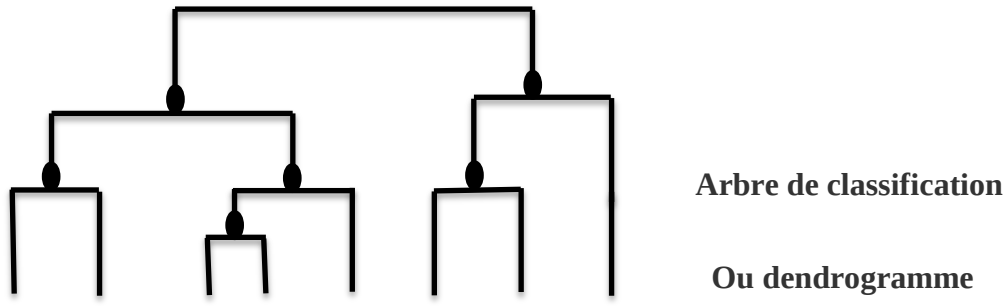



**Figure I.3** : regroupement et classification selon la méthode hiérarchique

### I.6.3.4 Les approches de la classification hiérarchiques

La construction d'une classification hiérarchique peut se faire de deux façons :pour la première, à partir d'une matrice symétrique des similarités entre les individus, un algorithme agglomératif forme initialement de petites classes ne comprenant que des individus très semblables, puis, à partir de celles-ci, il construit des classes de moins en moins homogènes, jusqu'à obtenir .Ce mode de construction est appelé Classification Ascendante Hiérarchique (CAH). Le second mode de construction d'une classification hiérarchique inverse le processus précédent. Il repose sur un algorithme divisif muni d'un critère de division d'un sous-ensemble de variables, et qui

procède par dichotomie successives de l'ensemble des individus tout entier, jusqu'à un niveau qui vérifient certaines règles d'arrêt et dont les éléments constituent une partition de l'ensemble des individus à classer. Ce mode de construction s'appelle la Classification Descendante Hiérarchique, et bien il y a deux approches comme suis :

- **Ascendantes :**

Pour Classifier des données quel que soit ses natures, on va dire que c'est regrouper entre eux des objets similaires selon tel ou tel critère ; pour un niveau de précision donné, deux individus peuvent être confondus dans un même groupe, alors qu'à un niveau de précision plus élevé, ils seront distingués et appartiendront à deux sous-groupes différents.

D'un autre point de vu ; Etant donné un ensemble d'observations, une **hiérarchie** sur cet ensemble est une collection de groupes d'observations (clusters) tels que : L'ensemble complet des données est un cluster. Chacune des observations est un cluster (singleton). Etant donnés deux clusters de la hiérarchie, ou bien ils n'ont aucune observation en commun, ou bien l'un est inclus dans l'autre (pas de chevauchement). Pour des raisons pratiques, il est commode d'imposer également que chaque cluster (excepté les singletons) est partitionné en exactement deux clusters de la hiérarchie. Une telle structure peut se représenter par un "dendrogramme" (ou "arbre").

- **Descendantes :**

Les méthodes de classification descendante hiérarchique sont itératives et procèdent à chaque itération au choix du segment de l'arbre hiérarchique à diviser, et au partitionnement de ce segment. La différence entre ces méthodes, développées jusqu'à présent dans la littérature, figure dans les critères qu'elles utilisent pour choisir le segment à diviser ainsi dans la manière dont elles divisent le segment. Le choix de tels critères dépend généralement de la nature des variables caractérisant les individus à classer.

### I.6.4 Méthode de classification ascendante hiérarchique

#### I.6.4.1 Définition

La classification ascendante hiérarchique est l'idée de créer à chaque étape une partition obtenue en agrégeant 2 à 2 les éléments représenté sous forme des vecteurs les plus proches en savant que un **Elément** est un individus ou groupe d'individus donc par principe chaque point ou individus ou cluster est progressivement « absorbé » par le cluster le plus proche.[10]

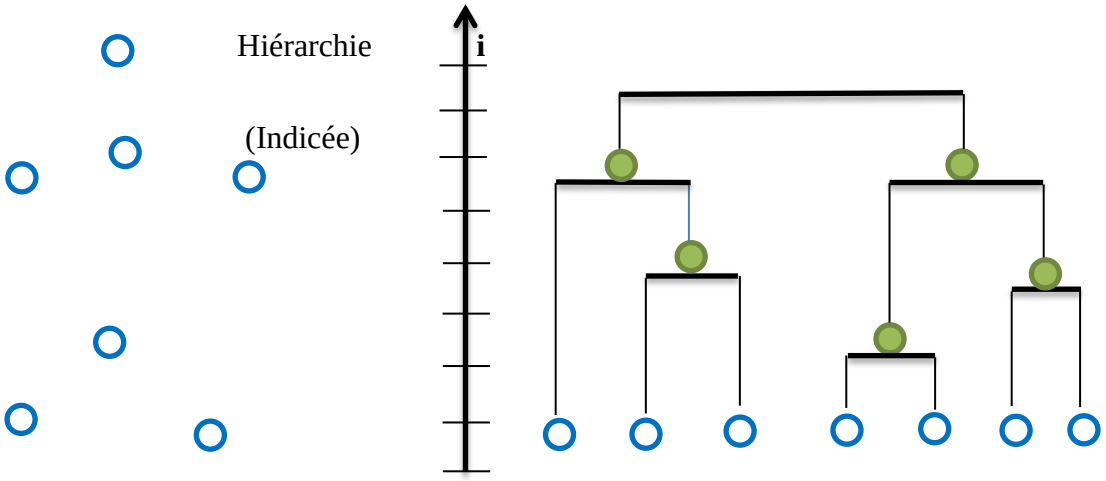



**Figure I.4** : exemple de Bissons 2001

#### I.6.4.2 Principe d'algorithme

La classification ascendante hiérarchique (CAH) est une méthode de classification itérative dont le principe est simple :

i. On commence par calculer la dissimilarité entre les N objets.
ii. Puis on regroupe les deux objets dont le regroupement minimise un critère d'agrégation donné, créant ainsi une classe comprenant ces deux objets.
iii. On calcule ensuite la dissimilarité entre cette classe et les N-2 autres objets en utilisant le critère d'agrégation. Puis on regroupe les deux objets ou classes d'objets dont le regroupement minimise le critère d'agrégation.

iv. On continue ainsi jusqu'à ce que tous les objets soient regroupés.

Ces regroupements successifs produisent un arbre binaire de classification (dendrogramme), dont la racine correspond à la classe regroupant l'ensemble des individus. Ce dendrogramme représente une hiérarchie de partitions. On peut alors choisir une partition en tronquant l'arbre à un niveau donné, le niveau dépendant soit des contraintes de l'utilisateur (l'utilisateur sait combien de classes il veut obtenir), soit de critères plus objectifs.

### I.6.4.3 Les Avantages

La (CAH) est une méthode de classification qui présente les avantages suivants :

- ✓ On travaille à partir des dissimilarités entre les objets que l'on veut regrouper. On peut donc choisir un type de dissimilarité adapté au sujet étudié et à la nature des données.
- ✓ L'un des résultats est le dendrogramme, qui permet de visualiser le regroupement progressif des données. On peut alors se faire une idée d'un nombre adéquat de classes dans lesquelles les données peuvent être regroupées.
- ✓ L'algorithme du « CHA » en générale consiste à fournir un ensemble de partitions de moins en moins fines obtenues par regroupement successifs de parties.

Donc le schéma de cet algorithme est le suivant :

A. Les classes initiales sont les individus eux-mêmes.
B. On calcule les distances entre les classes.
C. Les deux classes les plus proches sont fusionnées et remplacées par une seule.
D. Le processus reprend en B jusqu'à n'avoir plus qu'une seul classe, qui contient toutes les observations. »

On peut donnez l'algorithme d'une manière simple comme suit :

- Initialisation : les classes initiales sont les n singletons individus. Calculer a matrice de leurs distances deux à deux,
- Itérer les deux étapes suivantes jusqu'à l'agrégation en une seule classe :

✓ Regrouper les deux éléments (classes) les plus proches au dens distances entre groupes choisie

✓ Mettre à jour le tableau de distances en remplaçant les deux classes regroupées par la nouvelle et en calculant sa distance avec chacune des autres classes,

✓ Nécessite de définir une distance entre groupes d'individus (appelé stratégie d'agrégation),

✓ Nécessite de choisir le nombre de classe à retenir.

Alors, « l'algorithme agglomératif fonctionne donc en recherchant à chaque étape les classes les plus proches pour les fusionner, et l'étape la plus importante dans l'algorithme réside dans le choix de la distance entre deux classes. Les algorithmes les plus classiques définissent la distance entre deux classes à partir de la mesure de dissimilarité entre les objets constituant chaque groupe. De nombreuses distances sont ainsi possibles ce qui veut dire multiples algorithmes :

**1. L'algorithme Single-Linkage** où la distance entre deux clusters est représentée par la distance minimum entre toutes les paires de données entre les deux clusters (paire composé d'un élément de chaque cluster), nous parlons alors de saut minimum.

Le point fort de cette approche est qu'elle sait très bien détecter les classes allongées, mais son point faible est qu'elle est sensible à l'effet de chaîne[1], [Tufféry,2005] et donc moins adaptées pour détecter les classes sphérique.[7]

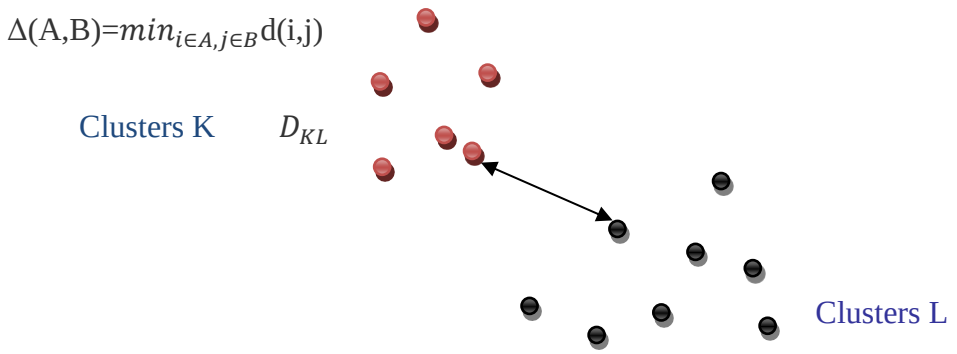



**Figure I. 5** : schéma de la classification « Single-Linkage »

[1] Nous appelons effet de chaine lorsque deux points très éloignés l'un de l'autre mais reliés par une suite de points très proche les uns des autres sont rassemblés dans la même classe.

**2. L'algorithme Complete-linkage** : où la distance entre de clusters est représentée par la distance maximum entre toutes les paires de données entre les deux clusters, nous parlons alors de saut maximum ou de critère du diamètre.

Par définition cette approche est très sensible aux point aberrants donc elle est peu utilisé [Tufféry, 2005]. Et bien on démontre aussi le résultat de cet algorithme et comment il fonctionne. [11]

La formule de la distance maximale est :

$$\Delta (A, B)=max_{i\in A, j\in B} d (i,j)$$

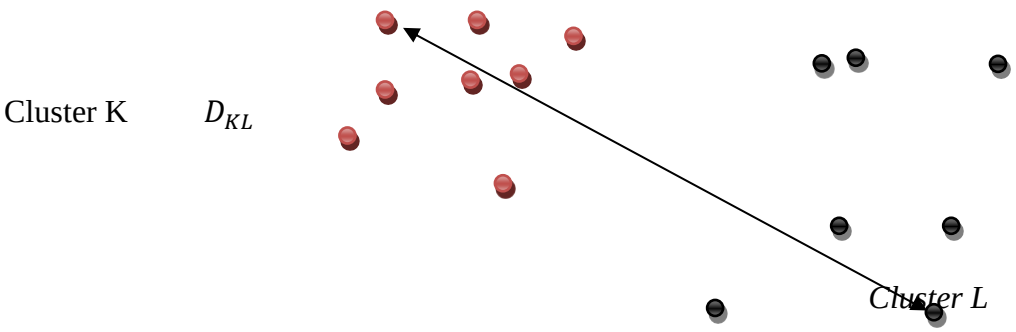



**Figure I.6 :** schéma de la classification « Complete-Linkage »

**3. L'algorithme Average-Linkage** : propose de calculer la distance entre deux clusters en prenant la valeur moyenne des distances entre tous les couples d'objets des deux clusters. Nous parlons aussi de saut moyen. Cette approche tend à produire des classes de même variance.

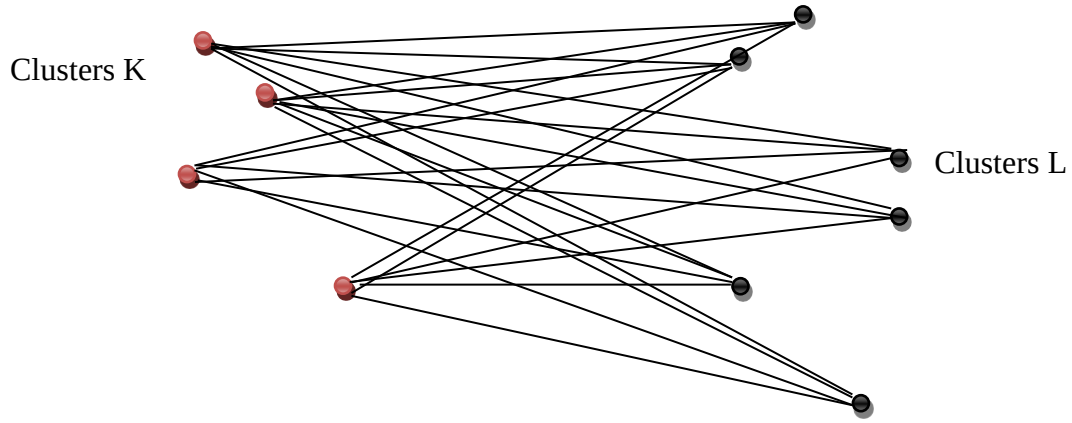



**Figure I.7 :** schéma de la classification «  Average-Linkage  »

**4. L'algorithme Centoid-linkage (**ou saut barycentrique) : définit, quant à lui, la distance entre deux clusters comme la distance entre leur centre de gravité. Une telle méthode est plus robuste aux points aberrants. Toutefois, elle est limitée aux données quantitatives numériques pour lesquelles le calcul du centre de gravité est possible. »

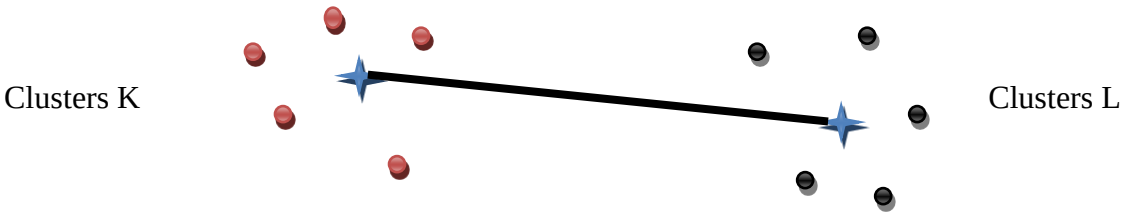



**Figure I.8 :** schéma de la classification «  Centoid-Linkage  »

### I.6.5 L'algorithme Ward-Linkage ou (méthode du saut Ward) :

#### I.6.5.1 Introduction

Si on peut considérer E comme un nuage d'u espace$\boldsymbol{R^P}$, on agrège les individus qui font le moins varier l'inertie intra-classe. A chaque pas, on cherche à obtenir un minimum local de l'inertie intra-classe ou un maximum de l'inertie interclasse.

L'indice de dissimilarité entre deux classes(ou niveau d'agrégation de ces deux classes) est alors égal à la perte d'inertie-classe résultant de leur regroupement ; Le regroupement des données soit des points ou autre démarre d'une façon hiérarchique aléatoire et finisse on donnant l'arbre avec ses différents niveaux. [13]

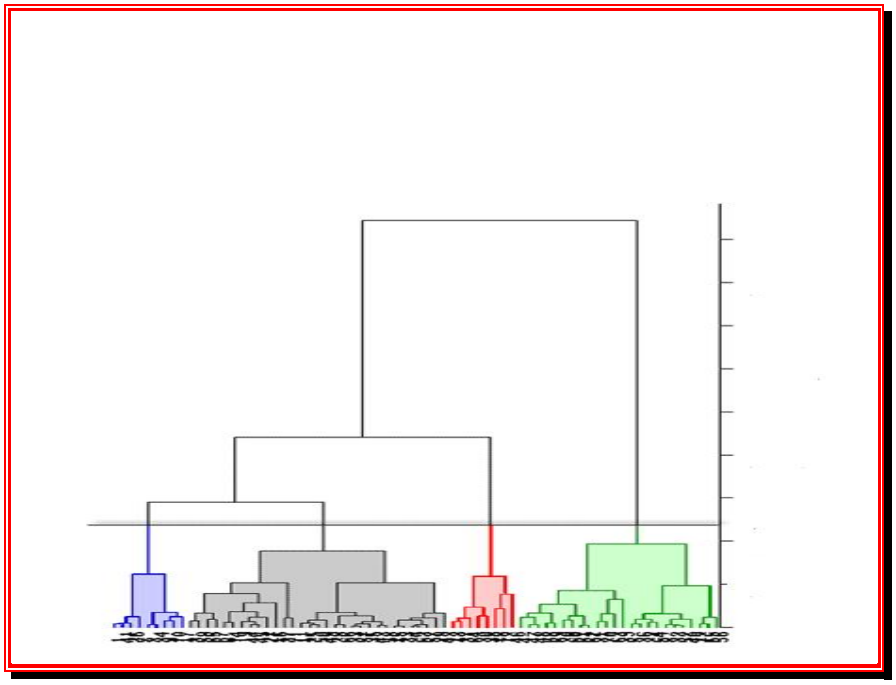


**Figure I.9** : classification selon la méthode du Ward

#### I.6.5.2 Définition

C'est la méthode la plus courante. Elle consiste à réunir les deux clusters dont le regroupement fera la moins baisser l'inertie interclasse. C'est la distance de Ward qui est utilisée pour : la

distance entre deux classes est celle de leurs barycentres au carré, pondéré par les effectifs des deux clusters. On suppose tout de même l'existence de distances euclidiennes.

Cette technique tend à regrouper les ensemble représentant les petites classes.

On calcule cette inertie entre les classes :

$G_A$ : Centre de gravité de la classe A (poids$P_A$).

$G_B$ : Centre de gravité de la classe B (poids$P_B$).

$G_{AB}$ : Centre de gravité de leur réunion.

$$G_{AB} = \frac{P_A G_A + P_B G_B}{P_A + P_B}$$

L'inertie interclasse étant la moyenne des carrés des distances des centres de gravité des classes au centre de gravité total, la variation d'inertie interclasse, lors du regroupement de A et B est égale à :

$P_A d^2(G_A, G) + P_B d^2(G_A, G) - (P_A + P_B)\, d^2(G_{AB}, g)$ elle vaut :

$$\Delta(A, B) = \frac{P_A P_B}{P_A + P_B}\, d^2(G_A, G_B)$$

A chaque itération, on agrège de manière à avoir un gain minimum d'inertie intra-classe Δ (A, B)=perte d'inertie intra-classe due à cette agrégation. [13]

#### I.6.5.3 Avantages de la méthode

Comme tous les autres algorithmes d'apprentissage la méthode de Ward représente des avantages sur le domaine d'application parmi ces derniers on a :

♦ L'algorithme peut produire un ordre des objets, qui peut être instructif pour l'affichage des données. Des grappes plus petites sont générées, ce qui peut être utile pour la découverte.

♦ Cette méthode est la plus appropriée pour les variables quantitatives, et non des variables binaires.

♦ Notons que l'interprétation de saut Ward est plus propre que l'interprétation obtenue plus tôt à partir de la méthode de liaison complète. Ceci suggère que la méthode de Ward peut être préférable pour des données actuelles.

♦ Elle est particulièrement adaptée lorsque la classification est effectuée après une analyse factorielle, les objets à classer étant repérés par leurs coordonnées sur les premiers axes factoriels (on obtient directement le carré de la distance en additionnant les carrés des coordonnées).De plus le critère d'inertie mis en œuvre dans la méthode de Ward la rend particulièrement compatible avec les analyses factorielles pour une utilisation complémentaire des approches.

♦ La méthode de Ward se joint à grappes afin de maximiser la probabilité à chaque niveau de la hiérarchie dans les hypothèses suivantes:

- Mélange normale multi variée.
- L'égalité des matrices de covariance sphériques.
- Probabilités d'échantillonnages égaux.
- Il faut noter aussi que cette méthode ne demande pas la fixation a priori du nombre de classes. [9]

#### I.6.5.4 Inconvénients de la méthode

Comme il y a des avantages il y a aussi des inconvénients alors ; La méthode de Ward a tendance à rejoindre les grappes avec un petit nombre d'observations, et il est fortement biaisé vers la production de grappes avec à peu près le même nombre d'observations. Il est également très sensible aux valeurs aberrantes (Milligan, 1980).

Quelque soient les critères utilisés, les distances et autres paramètres, il n'y aura jamais la certitude d'avoir atteint la meilleure Solution.

### I.6.5.5 Remarques sur la méthode

■ Avec la méthode de Ward, on agrège à chaque itération les classes dont l'agrégation fait le moins d'inertie interclasse il s'agit donc d'une optimisation pas-a-pas, qui ne dépend pas d'un choix initial arbitraire.

■ L'algorithme mis en œuvre scanne le fichier d'entrée de données et les stocke dans une liste chaînée. A chaque étape de clustering, la liste chaînée est modifié pour refléter le nombre et le contenu des clusters existants actuellement.

■ Ward (1963) décrit une classe de méthodes de classification hiérarchique, y compris la méthode de variance minimum.

### I.6.5.6 Principe

**L'algorithme de Ward**

**Entrée :** Une base d'exemples S

Une hiérarchie H contenant |S| clusters

**Sortie :** Une hiérarchie H mise à jours

**Etapes :** l'utilisateur va suivre ces étapes jusqu'à obtenir la sortie désirée

●Etablir la table TD (table des distances) des valeurs de D(x,y),x et y $\in$ S parcourant S.

●Tant que la table TD à plus d'une colonne faire

Choisir les deux sous-ensembles hi,hj de S tels que D(hi,hj) est la plus petite.

Supprimer Hj de la table, et remplacer hi par hi$\cup$hj (pour les colonnes et lignes).

Ajouter un nouveau cluster dans la hiérarchie H dont les fils sont hi et hj.

Calculer les distances de Ward entre hi∪hj et les autres éléments de la table. Et mettre à jour la table.

Fin tant que.

Retourner H

## I.7 Apprentissage Semi-Supervisé

### I.7.1 Introduction

Effectué de manière probabiliste ou non, généralement si on a des données et on est toujours dans le terme d'apprentissage alors il vise à faire apparaître la distribution sous-jacente des « exemples » dans leur espace de description. Il est mis en œuvre quand des données (ou « étiquettes ») manquent… Le modèle doit utiliser des exemples non-étiquetés pouvant néanmoins renseigner.

### I.7.2 Définition

C'est une classe de techniques d'apprentissage automatique qui utilise un ensemble de données étiquetées et non-étiquetés. Il se situe ainsi entre l'apprentissage supervisé qui n'utilise que des données étiquetées et l'apprentissage non-supervisé qui n'utilise que des données non-étiquetées. Il a été démontré que l'utilisation de données non-étiquetées, en combinaison avec des données étiquetées, permet d'améliorer significativement la qualité de l'apprentissage. Un autre intérêt provient du fait que l'étiquetage de données nécessite l'intervention d'un utilisateur humain. Lorsque les jeux de données deviennent très grands, cette opération peut s'avérer fastidieuse. Dans ce cas, l'apprentissage semi-supervisé, qui ne nécessite que quelques étiquettes, revêt un intérêt pratique évident.

Un exemple d'apprentissage semi-supervisé est le Co-apprentissage, dans lequel deux classificateurs apprennent un ensemble de données, mais en utilisant chacun un ensemble de caractéristiques différentes, idéalement indépendantes. Si les données sont des individus à classer en hommes et femmes, l'un pourra utiliser la taille et l'autre la pilosité par exemple.

En effet, étiqueter un échantillon (données, textes, pages WEB,...) est une opération coûteuse car elle nécessite un « expert ». [4]

### I.7.3 Problème

Avec en entrée un **petit** échantillon d'exemples étiquetés et un **grand** échantillon d'exemples non étiquetés tirés selon une loi fixée mais inconnue, il s'agit de trouver une procédure de classification de X dans {-1, +1}.

### I.7.4 Objectif

Minimiser l'erreur de classification qui est la probabilité qu'un exemple tiré aléatoirement soit mal classé par la procédure générer.

### I.7.5 Propriétés

**L**'apprentissage semi-supervisé pour améliorer les performances en combinant les données avec labels (peu) et sans labels (beaucoup).

**L**a classification semi-supervisé : entrainer sur des donnés avec labels et exploiter les données (beaucoup) sans labels.

**C**lustering semi-supervisé : clustering des donnés sans labels en s'aidants des donnés avec labels.

**H**ypothèse de base pour la plupart des algorithmes d'apprentissage Semi-Supervisé :

- Points proches ont probablement le même label.
- Deux points qui sont connectés par un chemin traversant des régions de forte densité doivent avoir le même label.

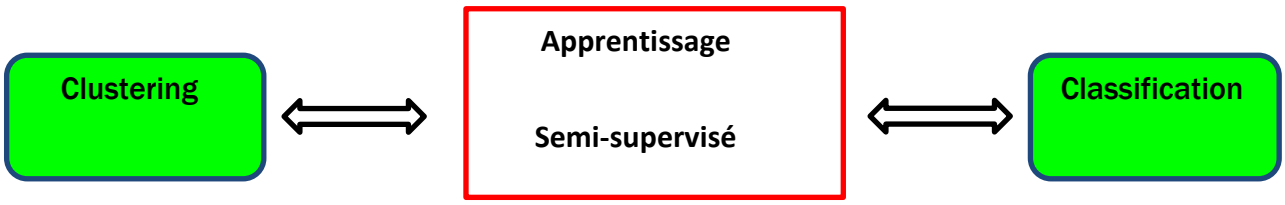



**Figure I.10 :** relations d'apprentissage semi-supervisé

### I.7.6 Quelques algorithmes d'apprentissage semi-supervisé

- EM Semi-Supervisé.
- Co-training.
- Traductive SVM's.
- Algorithmes à base de graphes.

## I.8 Conclusion

Dans ce chapitre nous avons présenté la notion d'apprentissage automatique et ses différents types ; nous nous sommes intéressé principalement par les méthodes clustering hiérarchique (méthode non supervisé).

## II.1 Introduction

Le clustering peut être considéré comme le plus important problème d'apprentissage non supervisé, de même, comme tous les autres problèmes de ce genre, l'algorithme du saut Ward tente de trouver une structure dans un ensemble de données hiérarchiques. Dans le cadre d'un processus qui regroupe des points par binômes et des clusters initialisés contenant des membres ; tout est basé principalement sur les distances entre les centres des clusters en accréditant sur la minimisation entre les centres. L'algorithme Ward calcule les centres de gravités au carré de chaque groupe qui dépend des effectifs du groupe « l'inertie intra-classe » ; puis le quotient entre la somme et la multiplication et la somme entre les poids de chaque groupe.

L'algorithme du Saut Ward affecte deux éléments donné à un groupe et par itérations le centre de ce groupe a son tour est affecté avec un autre élément a un autre groupe et ainsi de suite.

Donc l'algorithme Ward-Linkage favorise les hiérarchies équilibrés sauf et n'accepte que la certitude et la préférence dans la hiérarchie obtenu.

## II.2 Conception

### II.2.1 Organigramme d'algorithme de clustering à base de Ward

D'un point de vu générale l'algorithme consiste à classifier les points via une façon hiérarchique.

L'entrée de l'algorithme est une base d'exemples avec une hiérarchie initial des clusters des exemples .Quand les exemples (points) sont donnés l'algorithme va calculer la table des distances.

L'étape suivante ou il va calculer l'inertie minimale entre les centres de gravitée de chaque point représenté comme vecteur dans la TD initial et les grouper selon ce critère puis inclure ce groupe de deux dans la hiérarchie.

Et effet l'algorithme utilise à chaque fois les points (les exemples en entrée) d'après la Table des Distances alors cette dernière va être mise a jours ainsi que la hiérarchie.

L'algorithme va s'arrêter le faite qu'il va rester dans la Table des Distances q'un vecteur qui représente un point puisque il faut avoir au plus un  vecteur.

La hiérarchie finale est obtenue sous forme d'arbre.

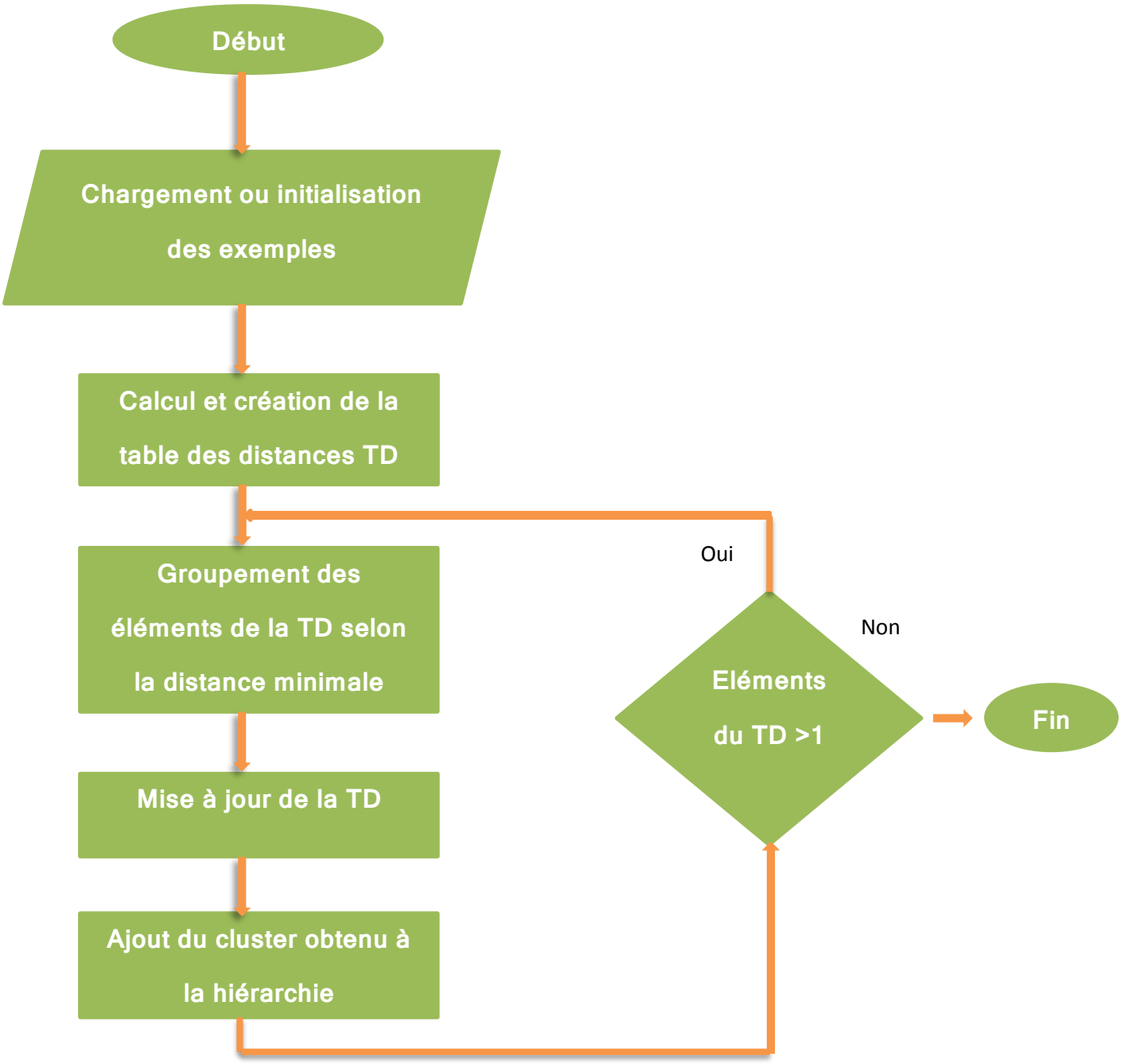



**Figure II.1 :** Organigramme de l'algorithme de Ward

Nous notons que notre base d'exemples contient plusieurs instances, chaque instance est un vecteur de Cinq dimensions (05 variables), chaque variable est générée de façon aléatoire selon des intervalles prédéfinis. Voir le tableau suivant :

| Variables | valeurs |
|---|---|
| variable 1 | -300..0 |
| variable 2 | 0.. 5 |
| variable 3 | -30..0 |
| variable 4 | -1.. 0 |
| variable 5 | -1.. 0 |

L'objectif de l'algorithme est de fournir une hiérarchie de nœuds, chacun d'eux contient les exemples appartenant au même cluster.

### II.2.2 Diagramme Use Case

Le mode de  fonctionnement qui va être suivi par l'utilisateur d'un point de vue informel afin d'obtenir le regroupement et l'ensemble des clusters est comme suis :

✓ Préparation des points « étant des vecteurs »

- Charger les points d'entrée.
- Générer les points.

✓ Clustering hiérarchique ascendant

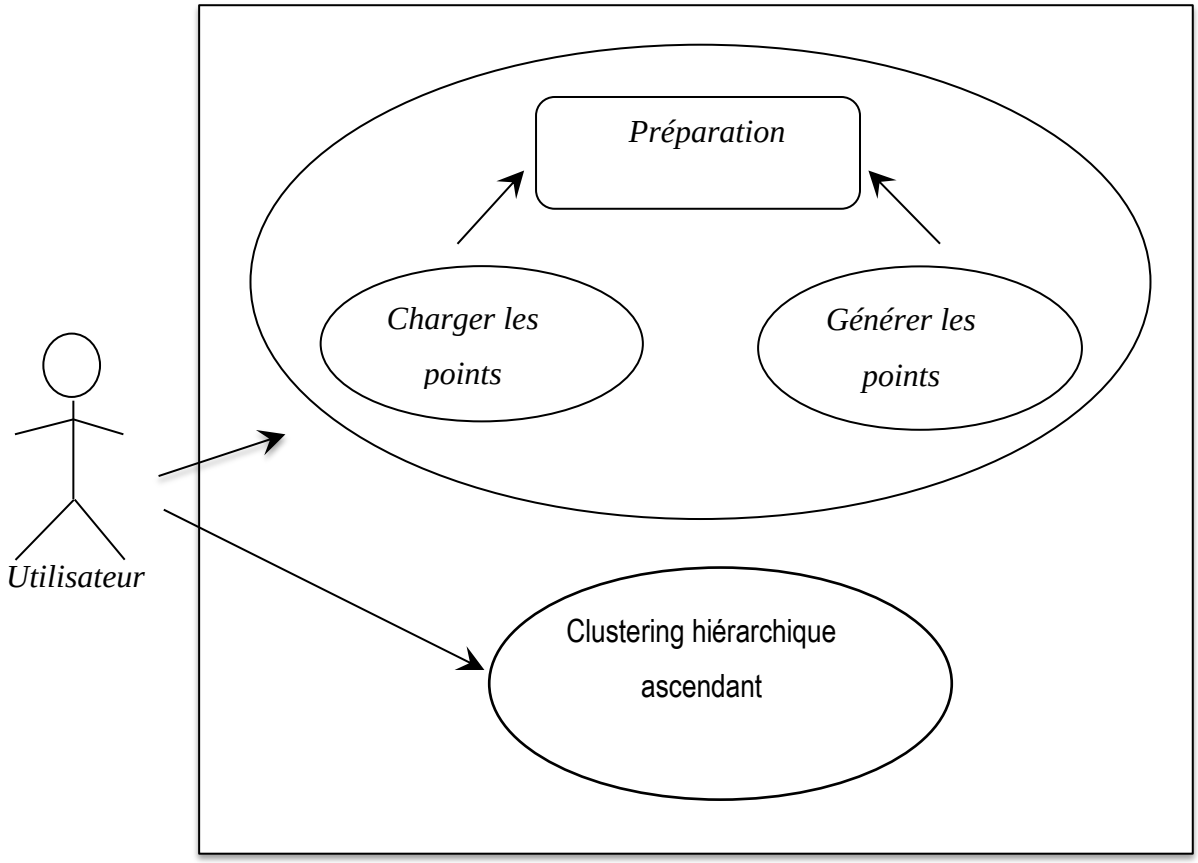



**Figure II.2 :** Diagramme de cas d'utilisation pour l'algorithme de Ward

### II.2.3 Diagramme de séquence hiérarchique

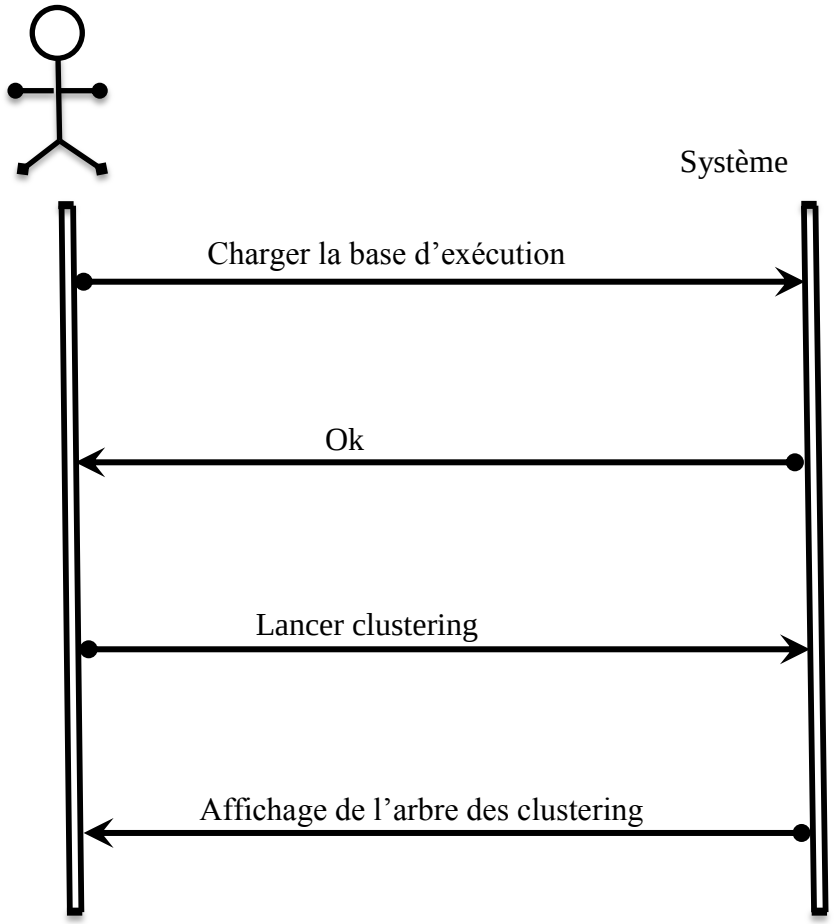
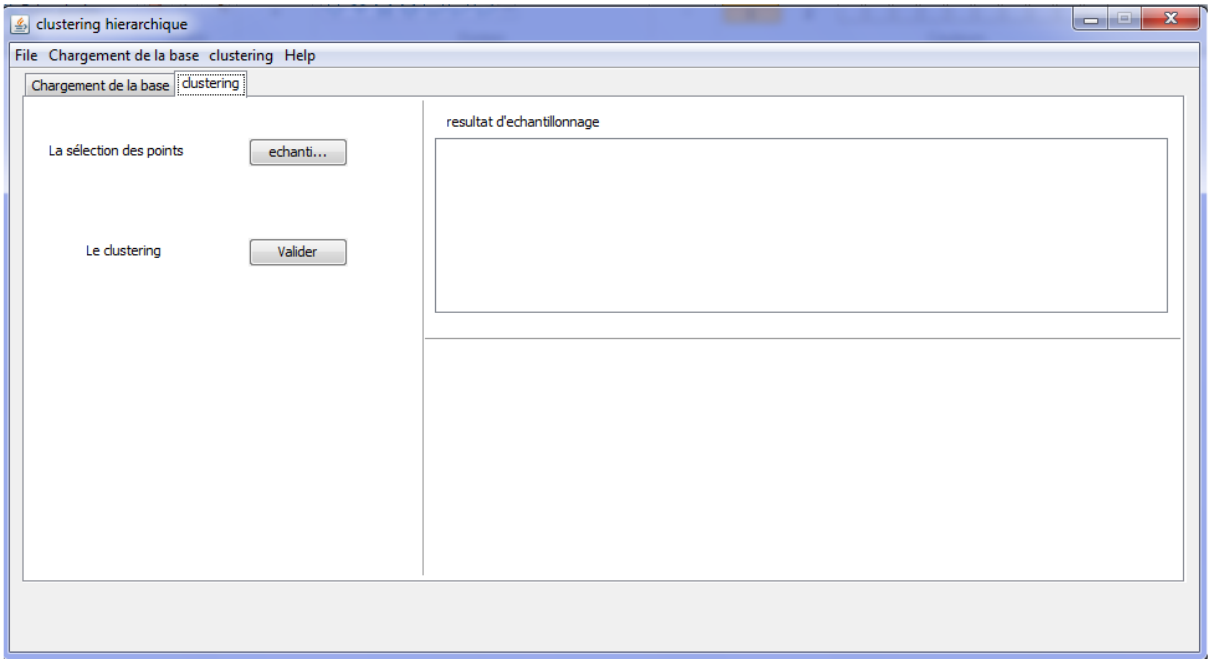


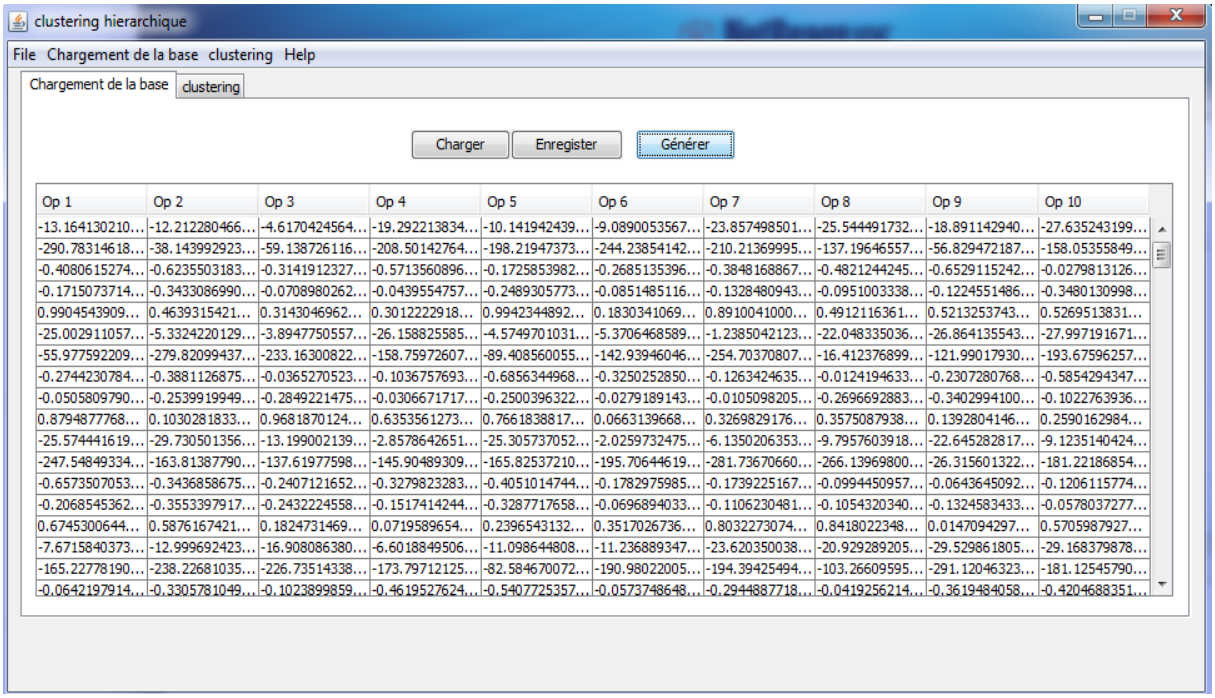

**Figure II.3 :** Diagramme de séquence

## II.3 Implémentation

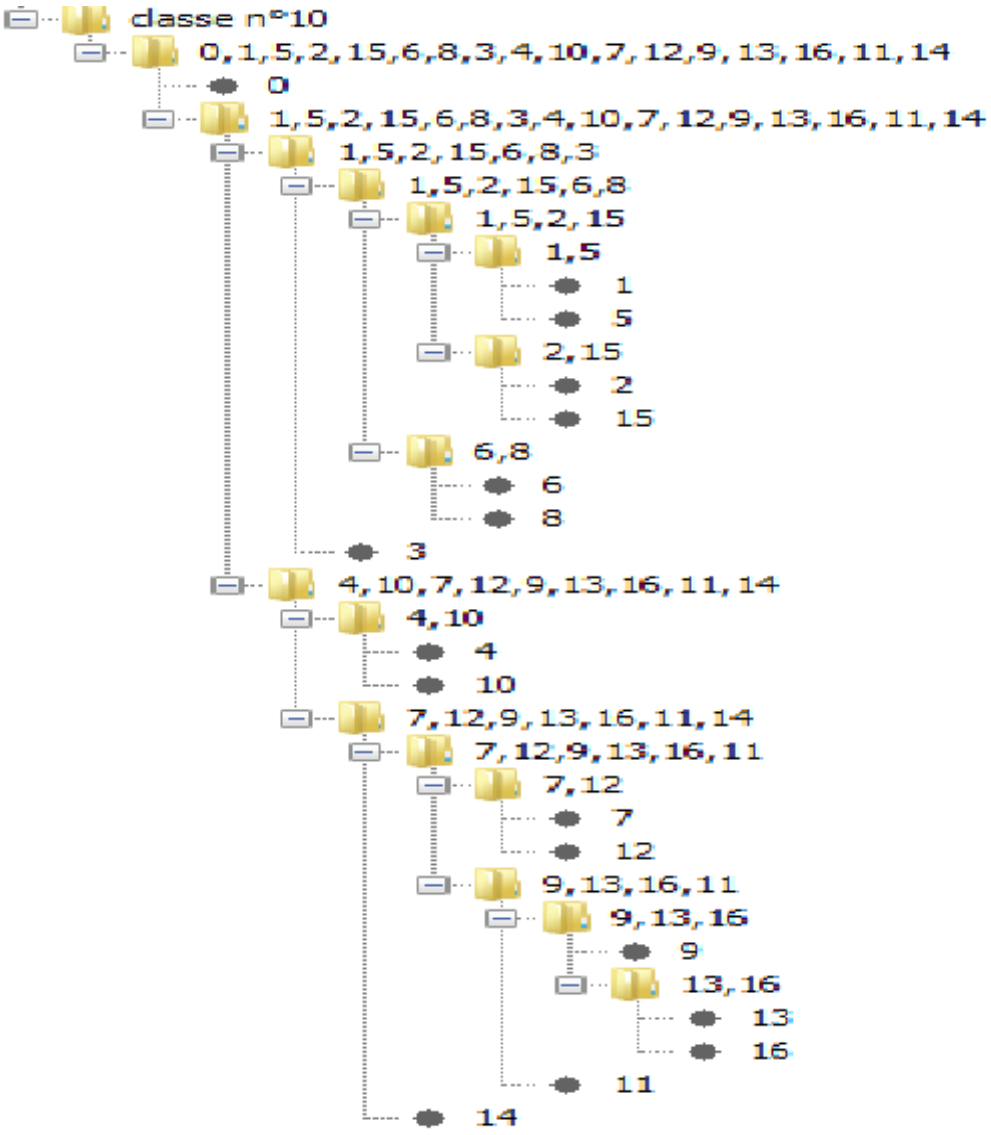
La figure suivante montre notre fenêtre principale




La fenêtre suivante permet le chargement des exemples



La figure suivante montre le résultat de clustering ascendant pour un jeu d'exemples




## II.4 Conclusion

Dans ce chapitre on a présenté une approche hiérarchique basons sur l'algorithme de Saut Ward ou on démarre par un ensemble d'éléments, par la suite les grouper dans des clusters hiérarchiques de deux nombre seulement .L'algorithme regroupe les ensembles/ points selon le mode ascendant de en utilisant l'inertie. (Qui doit être minimisée pour chaque cluster).

# Conclusion générale

Dans ce mémoire on a présenté une version de l'algorithme hiérarchique ascendant ; c'est le l'algorithme de Ward qui permet de regrouper les individus dans un ensemble des clusters sous forme d'une hiérarchie. On observe que l'arbre construit est composé de plusieurs partitions qui ne changent pas tant que les données sont toujours les mêmes donc l'algorithme repose sur la notion de la stabilité.

Comme perspectives nous pouvons implémenter d'autres algorithmes de clustering, tels que les cartes de kohonen, l'expectation –maximisation, …etc. Nous considérons que Ward reste toujours la meilleure parmi les autres algorithmes de la classification hiérarchique et la préférable à utiliser.

Au niveau des difficultés, nous sommes confrontés à des problèmes de compréhension. Comme la plupart des documents qu'on pouvait nous prêter n'était pas nets et lucides ;
Malgré les nombreuses heures passées devant l'ordinateur, les déplacements, les différents documents à consulter, etc…nous avons réellement eu beaucoup de plaisir à faire ce travail, et choisir ce thème.

# Référence

**[1]** http://fr.wikipedia.org/wiki/Apprentissage_automatique#cite_ref-4.

**[2]** Christian Gagné .apprentissage et reconnaissace- GIF- 4101/GIF-7005 .Université Laval Quebec Canada.2010

**[3]** http://fr.wikipedia.org/wiki/Apprentissage_automatique.

**[4]** Celeux G,.Diday E., Govaert G., Lechevallier Y., Ralam-Bondrainy H. Classification Automatique des Données. Bordas, Paris, 1989.

**[5]** Haytham ELGHAZEL, préparée au sein du laboratoire LIESP – Universit´e Claude Bernard Lyon 1, d´ecembre 2007.

**[6]** http://www.deenov.com/Data/Sites/1/docs/classification-tutoriel.pdf . date du dernier accès 15/06/13

**[7]** Blum, A., Mitchell, T. combining labeled and unlabeled data with co-training. COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann, 1998, p. 92-100.

**[8]** KDX: An Indexer for Support Vector Machines Navneet Panda, Edward Y. Chang, TKDE.2006.101

**[9] (**fr) Gilbert Saporta, *Probabilités, Analyse des données et Statistiques*, Paris, Editions Technip, 2006, 622.

**[10]** Valery Ridde et Christian Dagenais, les Presses de l'Université de Montreal.

**[11]** http : //www.asu.edu/~jye02.

**[12]**http://www.xlstat.com/fr/produits-solutions/fonctionnalite/classification-ascendante-hierarchique-cah.html.

**[13]**Pierre-Luis GONZALEZ, méthode de classification, 2008.

# Résumé

Le clustering est considéré comme l'une des facultés facinantes du cerveau humain,plusieurs implémentations artificielles ,sont proposées pour simuler ce processus,nous distinguons le clustering à base de centroides,le clustering hiérarchique,le clustering à base de distibution…etc.

Dans ce travail nous avons choisi la classe de clustering hiérarchique , et en particulier nous avons conçu et implémenter l'algorithme ascendant de Ward.Ce dernier possède beaucoup d'avantages, par rapport à d'autres algorithmes tels que K-means.



# Abstract

The clustering process is considered as the most fascinating capabilities of the human brain,in fact ,several implementations have been proposed to simulate this process on machines ,we distinguish centroide clustering,hiérarchique clustering…

In this work we propose a hierarchical clustering method based on Ward algorithm ,this later is very stable in comparison with other algorithms such as K-means.