REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mouloud MAMMERI de Tizi-Ouzou

Faculté de Génie Electrique et d'Informatique
Département d'Automatique

PROJET DE FIN D'ETUDES

En vue de l'obtention du diplôme

D'INGENIEUR D'ETAT EN AUTOMATIQUE

Thème

fication

Proposé par :
Mme. Boudjemaa F

Présenté par :
Mr. Ramdani Smaïl
Mr. Boukhari Sofiane

Soutenu le : 15 / 07 /2009 Devant le jury d'examen composé de :

Mr. HAMOUCHE K
Mlle. CHILLALI O
Mme. ALKAMA S

Promotion 2009

Remerciement

Nous tenons à remercier notre promotrice Mme BOUDJEMAA de nous avoir proposé ce sujet et nous avoir consacré un temps précieux.

Nous lui exprimons notre sincère reconnaissance pour son aide et ses conseils tout au long de ce travail.

Nous remercions également les membres de jury d'avoir accepté d'examiner notre travail.

Mes remerciements à mes amis.

Dédiasses

Je dédie ce modeste travaille à :

Mes chers parents qui m'ont soutenu tout au long de mon cursus,

À mes frères ,

À toute la famille BOUKHARI et MERBOUTI,

À tous mes amis.

Sommaire

Introduction générale

Introduction générale

La classification est une activité mentale qui intervient fréquemment dans la vie courante.

En effet, les objets, quelque soit leur nature, sont souvent répertoriés par rapport à des catégories ou des classes auxquelles ils sont censés appartenir.

De nos jours , la classification est une démarche qui est appliquée dans un nombre important de domaines et les termes utilisés dépendent de la discipline scientifique ou du domaine d'application , on parle de classification en reconnaissance des formes , de discrimination ou de prédiction en statistique , d'apprentissage de concept ou d'apprentissage inductif en apprentissage automatique ,aussi dans le marketing on parle de typologie .

La classification a pour but de classer en un nombre fini d'objets, de tel sorte que les objets appartenant à une même classe soient plus semblable que ceux appartenant à des classes différentes.

Cette démarche est relativement difficile à formaliser surtout quand on se place dans un contexte non supervisé, c'est à dire quand on ne dispose d'aucune information à priori sur la structure de l'ensemble des objets à classer ni le nombre exacte de classes.

La classification automatique fait actuellement l'objet de recherche poussées, chose qui a donné comme fruit une multiplicité de méthodes de classification automatique, cette multiplicité rend souvent l'utilisateur inerte devant le chois de la méthode qui lui convient selon le domaine d'application ou selon la taille du fichier de donnée.

Ce travaille consiste en l'étude comparative de méthode de classification afin d'en tirer les avantages et les inconvénients d'une méthode par rapport à une autre dans le but de faciliter la tache du chois de la méthode adéquate

Dans le premier chapitre on va exposer d'une manière générale les différentes méthodes de classification qui existe en ce moment .Après dans le second chapitre on va détailler

L’algorithme de quelque méthode qui sont en l’occurrence la méthode de classification hiérarchique ascendante, la méthode des K-means, la méthode d’agrégation autour des centres mobiles et enfin la méthode ISODATA.

Dans le dernier chapitre on va tester les programmes des méthodes sous MATLAB en utilisant des fichiers de donnée artificielle.

Chapitre I

Méthodes de classification

I.1 Introduction

La classification automatique est une étape de base dans plusieurs disciplines, il s'agit d'une démarche très courante qui permet de mieux comprendre l'ensemble de données à analyser, elle tend à trouver une certaine structure dans une collection de données non étiquetées.

La classification automatique est un processus qui permet d'organiser un ensemble de données en classes cohérentes ou homogènes telle que, dans une même classe, les données présentant les mêmes caractéristiques de manière à ce que l'intersection, deux à deux, des classes formées donne un ensemble vide et que l'union de toutes les classes donne l'ensemble initial des données. Elle s'applique, a priori, sur n'importe quel type de données : tableau de contingence, tableau de distances, etc. La plupart des cas se déroule en trois étapes, et à l'aide de quelques paramètres indispensables : mesure de ressemblance, structure de la classification et type d'algorithme [1]. Ainsi, étant donné un ensemble $X=\{x_1, x_2,\ldots, x_m\}$ d'un certain nombre *m* d'objets x_i, i=1,2, …,.*m*, où chaque objet x_i est caractérisé par n attributs correspondant à des mesures effectuées sur l'ensemble des objets. Il s'agit, donc, de partitionner l'ensemble X en un nombre K classes c_k , k=1,2,…,K. Pour réaliser une classification automatique il existe plusieurs méthodes que l'on peut distinguer, tout d'abord, selon que l'on dispose de toutes les informations *a priori* sur les données. Dans ce cas, la classification est dite supervisée. Dans le cas où l'information n'est pas complète ou sans information *a priori*, on parlera de classification non- supervisée ou clustering.

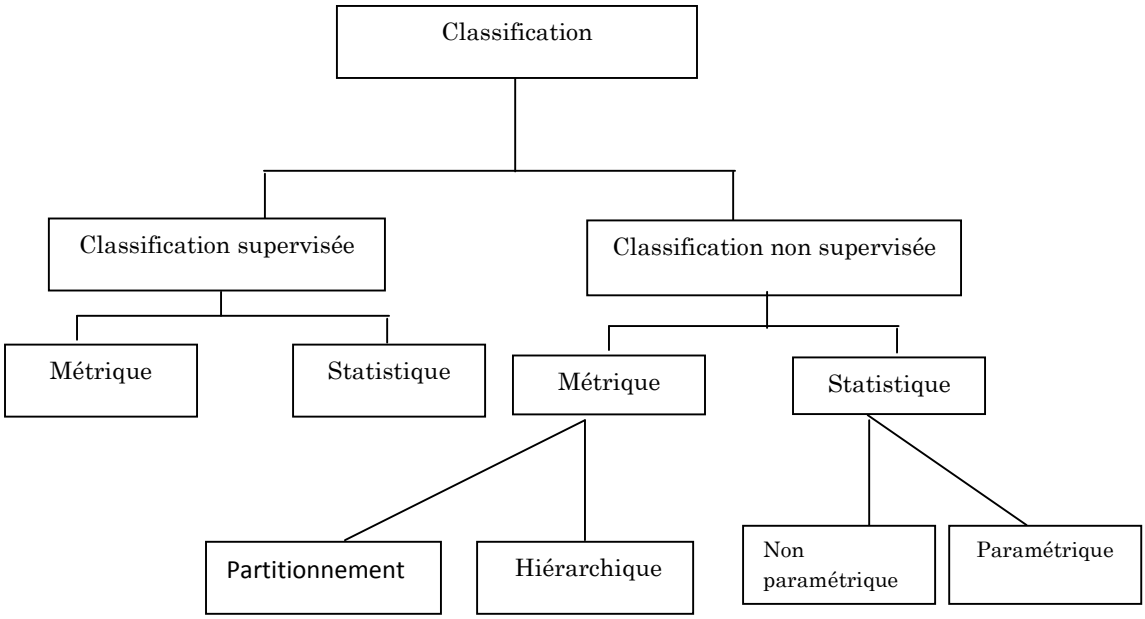


Fig. 1.1. Différentes méthodes de classification automatique

I.2 Classification supervisée

Ces techniques de classification sont des méthodes très étudiées. Leur objectif est de prédire la classe d'une nouvelle observation. La construction d'un classifieur peut être réalisée par deux approches différentes métrique et statistique.

Les méthodes métriques est un ensemble de techniques qui se basent sur l'utilisation de la notion de la fonction discriminante [2], donc. Pour un problème à K classes, on définit K fonctions discriminantes $d_k(x)$, q=1,2…K. La comparaison des valeurs prises par les fonctions discriminantes pour la valeur X d'une observation permet de classer l'objet correspondant. Plus précisément, l'observation X est assignée a la classe C_s si seulement si : $d_s(X) > d_k(X)$, $\forall k \neq s$

Les points de l'espace des attributs, pour lesquels les fonctions discriminantes $d_s(X)$ et $d_k(X)$ sont égales, sont situés sur une surface de dimension N-1 appelée *surface de décision* et définie par : $ds(X) = d_k(X)$, Cette surface partage l'espace en deux régions, l'une pour laquelle $d_s(X) > d_k(X)$ l'autre pour laquelle $d_s(X) < d_k(X)$.

On ne peut parler des méthodes statistique que lorsque on fait explicitement appel aux caractéristiques statistiques de la distribution des observations ou la notion de fonction

de densité de probabilité (fdp) tel que : Une observation est considérée comme une réalisation particulière d'un vecteur aléatoire continu dont la distribution est définie par une fdp multi variables caractéristique de la classe à laquelle appartient l'observation. Ces méthodes utilisent la théorie de la décision qui constitue une approche statistique fondamentale pour résoudre les problèmes de classement et de la classification qui se trouvent posés en terme probabiliste, cette théorie permet d'effectuer un classement optimal à partir de la règle de Bayes basée sur la connaissance des probabilités à priori de chaque classe

La règle de décision de bayes consiste à choisir d'affecter l'individu à la classe dont la probabilité *a posteriori* (calcul par la formule de Bayes) est la plus grande et cette décision minimise le risque d'erreur de classification.

La formule de bayes prend en considération la probabilité a priori d'apparition des individus des différentes classes et de leurs distribution dans l'espace des descripteurs. La formule de bayes s'exprime comme suit :

$$P\left(C_K / x\right) = \frac{P_{rk} \cdot f_k(x)}{\sum_{k=1}^{K} P_{rk} \cdot f_k(x)} \qquad (1.1)$$

Avec P (C_k/x) la probabilité a posteriori que l'individu de coordonnées x appartienne à la classe k, Pr_K la probabilité a priori que l'individu appartienne à la classe k, f_K(x) la densité de probabilité de x si la classe est k.

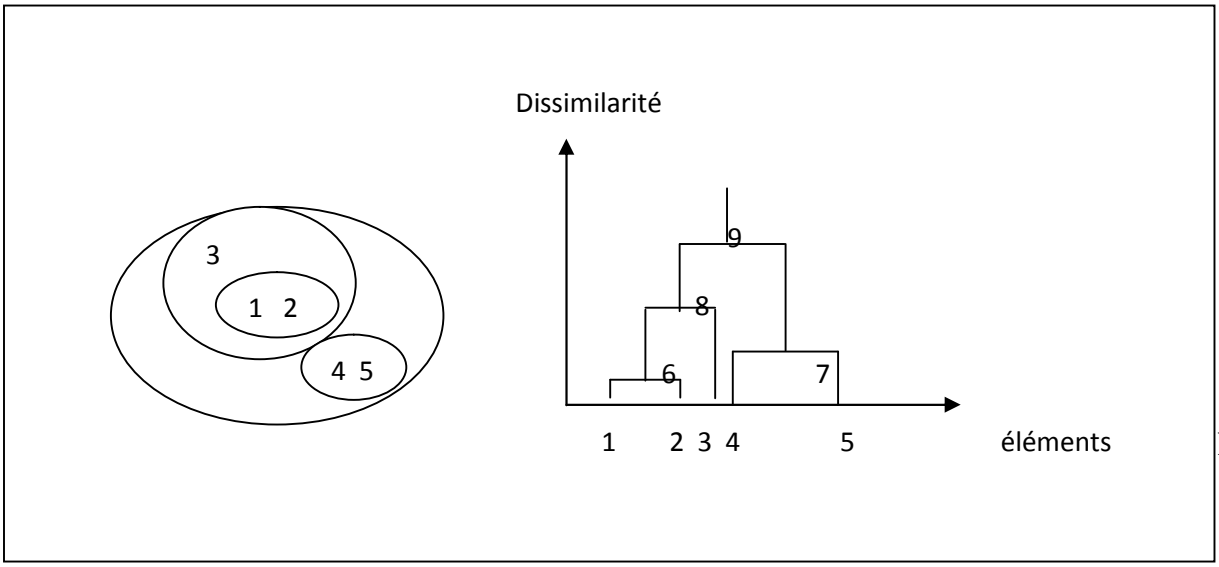
I.3 Classification non supervisée

Est une technique d'analyse des données utilisée dans plusieurs domaines, on ne dispose ni de nombre de classes ni de l'information sur l'appartenance des individus aux classes (prototypes) donc la classification revient à regrouper les observations ayant des caractéristiques similaires en classes distinctes en définissant une mesure de proximité entre les observations. Pour ce mode de classification on dispose de plusieurs techniques.

Parmi ces techniques on a les méthodes statistiques, les méthodes paramétriques et non paramétriques, les plus utilisées sont les méthodes métriques basées sur des notions mathématiques non probabilistes, elle consiste à montrer le degré de similarité entre objets en utilisant la notion de distance. On distingue deux principales catégories la première consiste à établir une hiérarchie de classe tandis que la seconde réalise un partitionnement de l'espace des observations.

I.3.1 La classification hiérarchique

La classification hiérarchique consiste à construire un arbre de classes appelé dendrogramme (figure 1.2), cette construction se fait par deux manières ascendantes [3] et descendantes [4]. La première considère chaque objet de l'ensemble de données comme des classes initiales, et à chaque étape on fusionne deux classes qui maximise (resp. minimise) un critère de similarité (resp. dissimilartité).



la

Fig. 1.2 Exemple de dendrogramme

L'algorithme fondamental de classification ascendante hiérarchique se déroule da al façon suivante :

Étape *1* : on dispose de *m* individus à classer ;

Étape *2* : on construit la matrice de distances entre les *m* éléments et l'on cherche les deux plus proches, que l'on agrège en un nouvel élément. On obtient une première partition à *m-1* classes

Étape *3*: on construit une nouvelle matrice des distances qui résultent de l'agrégation, en calculant les distances entre le nouvel élément et les éléments restants (les autres distances sont inchangées). On se trouve dans les mêmes conditions qu'à l'étape 1, mais avec seulement *(m*-1) éléments à classer et en ayant choisi un critère d'agrégation. On cherche de nouveau les deux éléments les plus proches, que l'on agrège. On obtient une deuxième partition avec *m-2* classes et qui englobe la première ;

Étape *m* : on calcule les nouvelles distances, et l'on réitère le processus jusqu'à n'avoir plus qu'un seul élément regroupant tous les objets et qui constitue la dernière partition.

Fig. 1.3 algorithme de la classification hiérarchique ascendante

A l'inverse, les méthodes descendantes partent d'une classe unique formée de tous les objets et le divisent petit à petit jusqu'à la satisfaction de critère d'arrêt. Il existe de nombreuses manières de définir une mesure de similarité ou de dissimilarité. Parmi les plus courantes, citons :

Critère du lien minimum (Single-Link) :

Où C_a, C_b deux classes, d est la distance entre les deux éléments x et y

$d(C_a, C_b)=\min \{d(x, y) \text{ avec } x \in C_a \text{ et } y \in C_b\}$

Critère du lien complet (Complete-Link) :

d (C_a, C_b)=max {d(x, y) avec x∈C_a et y∈C_b}

Critère de la distance moyenne (Average-Link) :

$$d(C_a, C_b) = \frac{\sum_{i=1}^{n_a}\sum_{j=1}^{n_b} d(x_i, x_j)}{n_a * n_b} \quad \text{Avec } x_i \in C_a \text{ et } x_j \in C_b \qquad (1.2)$$

Où *na* et *nb* désignent respectivement les effectifs des classes *Ca* et *Cb*

A partir de ces mesures de similarité ou de dissimilarité se découle les méthodes de classification hiérarchique suivantes : la méthode de plus proche voisin, la méthode de voisin le plus éloigné, Les méthodes intermédiaires.

I.3.2 La classification non hiérarchique (partitionnement)

Contrairement aux algorithmes précédents les algorithmes de partitionnement construisent directement une partition des données en *K* classes. Ces algorithmes génèrent une partition initiale puis cherchent à l'améliorer en réattribuant les données d'une classe à l'autre en calculant une fonction de distance ; une fois cette fonction de distance est définie, la procédure de partitionnement consiste à réduire au maximum la distance entre les objets et augmente au maximum la distance entre classes, et chaque classe est représente par son centroïde ou par un noyau. Plusieurs techniques existent, parmi elle les K-means, nuée dynamique, Isodata, agrégation autour des centres mobiles, ces méthodes se ressemble dans le principe général, mais différent dans la façon de procéder pour aboutir à une partition finale. Par la suit en détaillera les principes de quelque méthodes telle que :

- **La Methode des K means :**

Cet algorithme a été proposé par McQueen en 1967 [MAC-67], son principe est de partitionner l'ensemble des points en un ensemble de classes prédéterminées. Des points initiaux sont choisis aléatoirement pour constituer les centres des classes .les autres points sont alors assignés à la classe dont le centre est le plus proche .ensuite, on recalcule les centres des classes et on répète le même processus jusqu' à avoir une partition optimale.

Géométriquement ,cela revient à partager l'espace des points en K zones définies par les plans médiateurs des segments de droite reliant deux centres de deux classes différentes D'une manière plus générales, cet algorithme se déroule de la manière suivante :

Soit un ensemble X de m individus caractérisés par n variables, à partitionner en K classes $\{c_1,\ldots\ldots c_k,\ldots\ldots\ldots c_K\}$. L'algorithme est le suivant

1. On choisit aléatoirement K centres de classes $\{v_1, v_2\ldots\ldots\ldots v_k\}$.ces k centres induisent une première partition de l'ensemble des m individus en K classes $\{c_1,c_2\ \ldots\ldots\ldots c_k\}$.
2. chaque individu est affecté à la classe la plus proche. Par exemple un individu X i est affecté à la classe c_k s'il est plus proche de v_k que de tous les autres centres.
3. on calcule les K nouveau centres de classes à chaque fois qu'un individu change de classe.
4. si le critère d'arrêt est vérifié on termine, sinon aller a 2.

Fig1 .4 algorithme des K-means

Plusieurs critères d'arrêt peuvent être utilisés :

-deux itérations successives conduisent à la même partition.

-un nombre maximal d'itérations, fixé à priori, est atteint.

-les centres des classes ne changent pas entre deux itérations successives.

-minimisation d'une fonction coût.

- ***La méthode d'agrégation autour des centres mobiles***

L'algorithme peut être imputé principalement à Forgy [5], bien que de nombreux travaux (parfois antérieurs : Thorndike, 1953). Bien qu'elle ne fasse appel qu'à un formalisme limité et que son efficacité soit dans une large mesure attestée par les seuls résultats expérimentaux, la méthode de classification autour de centres mobiles est probablement la technique de partitionnement la mieux adaptée actuellement aux

vastes recueils de données. elle est utilisée aussi bien comme technique de description et d'analyse que comme technique de réduction, généralement en association avec des analyses factorielles et d'autres méthodes de classification. Cette méthode peut être considérée comme un cas particulier de techniques connu sous le nom nuées dynamique. Cette méthode consiste à calculer les centres de gravité des classes initialement fixées par l'utilisateur dans façon aléatoire, puis de réaffecter à chaque classe les objets les plus proches de son centre de gravité formant ainsi une nouvelle partition. L'opération de réaffectation est répétée jusqu'à ce que les objets ne change plus de place pendant l'opération de réaffectation.

Les étapes de l'algorithme

Soit un ensemble I de *m* individus à partitionner, caractérisés par *n* variables. On suppose que l'espace $\Re^n$ supportant les *m* points est muni d'une distance appropriée notée *d* (souvent distance euclidienne usuelle).On désire constituer au maximum K classes.

1) On détermine *K* centres provisoires de classes par tirage pseudo-aléatoire. Les *K* centres induisent une première partition P^0 de l'ensemble des individus I en *K* classes

i=0

2) Tant que le critère d'arrêt n'est pas vérifié faire

3) Calculer les centre de gravité des classes de la nouvelle partition P^i

4) affecter chaque objet à la classe dont le centre est plus proche

5) $P^{i+1} \leftarrow P^i$

6) Fin tant que

7) retourner P^{i+1}.

Fig1.4 algorithme d'agrégation autour des centres mobiles

Le processus se stabilise et l'algorithme s'arrête soit lorsque deux itérations successives conduisent à la même partition, soit encore parce qu'un nombre maximal

d'itérations a été fixé *a priori*. Généralement, la partition obtenue finalement dépend du choix initial des centres.

- ***La méthode des nuées dynamique***

Proposé par Diday en 1971[6], les classes ne sont pas caractérisées par un centre de gravité, mais par un certain nombre d'observations à classer, dénommé « étalons », qui constituent alors un « noyau » ayant pour certaines utilisations un meilleur pouvoir descriptif que des centres ponctuels.

Soit, C={C_1,C_2,C_3,…….C_K} est un ensemble de classe, f est une application qui associe à cet ensemble un ensemble de noyau N={N_1,N_2,N_3……N_K}, et g une application qui associe à un ensemble des points (noyau) l'ensemble des classes constituées des éléments les plus proches d'un noyau.

Enfin, on se donne un critère (à valeur positives) qui permet de mesurer l'adaptation d'une famille de noyaux à une partition :

$$H(C,N)=\sum_{k=1}^{K}\sum_{x_i\in C_{k_i}} d^2(x_i,N_k) \qquad (1.3)$$

Ou d est la distance euclidienne. Pour C donné, f associe la famille de noyaux N qui minimise H. g associe la partition qui minimise H.

- ***La méthode ISODATA***

Le terme ISODATA est l'abréviation d'Itérative Self-Organization Data Analysis Techniques, La dernière lettre À est ajoutée pour faciliter la prononciation de terme ISODATA.

L'algorithme ISODATA [7] construit une partition initiale et affecte un individu à la classe dont le centre est le plus proche. Ensuite, les classes sont recomposées en suivant un certain nombre de règles :

- L'éclatement d'une classe si cette classe a une dispersion maximale ($Disp_{max}$) dépassant un seuil donné (D_{max}).

  Si une classe j doit être éclatée, on construit deux centres de classe, $z_{j'}$ et $z_{j''}$ identique à z_j sauf pour la coordonnée k' qui maximise la quantité $\sqrt{\frac{1}{n_j}\sum_{i\in I_j}(x_i^k - z_j^k)^2}$ et dans ce cas

$$z_{j'}^{k'} \leftarrow z_j^{k'} + D_{\max}\sqrt{\frac{1}{n_j}\sum_{i\in I_j}(x_i^{k'} - z_j^{k'})^2} \qquad (1.4)$$

$$z_{j''}^{k'} \leftarrow z_j^{k'} - D_{\max}\sqrt{\frac{1}{n_j}\sum_{i\in I_j}(x_i^{k'} - z_j^{k'})^2} \qquad (1.5)$$

- Agglomération de deux classes j' et j'' si la distance entre leurs centres est inférieure à un seuil donné D_{min}
- Suppression d'une classe j si son effectif est inférieur à un seuil donné N_{min}

Voila quelque formule pour calculer les paramètres utilisés dans cette technique :

Pour la distance d utilisée d est la distance euclidienne.

La distance minimale et maximale entre deux groupes d'objets est :

$D_{max} = \max d(x_i, x_j)$ $\quad d_{min} = \min d(x_i, x_j)$

La distance moyenne entre deux groupes d'objets est

$$d_{mean} = \frac{2}{n(n-1)}\sum_{i\neq j} d(x_i, x_j) \qquad (1.6)$$

La dispersion de la classe j est définie de la façon suivante :

$$Disp_{\max}(j) = \max_{1\le k\le P}\left\{\sqrt{\frac{1}{n_j}\sum_{i\in I_j}(x_i^k - z_j^k)^2}\right\} \qquad (1.7)$$

où n_j est l'effectif de la classe j, I_j est l'ensemble des indices des objets de la classe j, $z_j=(z_1^1,....z_j)$ est le centre de gravité de la classe j ;

L'algorithme figure 1.4 résume le fonctionnement d'ISODATA. La principale difficulté que l'on rencontre avec ISODATA est qu'il faut déjà avoir une certaine connaissance des données pour fixer les paramètres, tel que $Disp_{max}$ fixant l'éclatement d'une classe en deux, ainsi que le paramètre D_{min} fixant le rapprochement de deux classes. Pour les seuils sont fixe par l'utilisateur.

1)-Affecter les n objets à c classes au hasard.

2)-Calculer les centres des classes.

3)-Affecter chaque objet à la classe dont le centre est le plus proche.

4)-Eliminer les classes comportant moins de N_{min} objets, les individus alors orphelins sont affectés aux classes dont les centres sont plus proches.

5)-Calculer les centres des classes.

6)-Si deux classes sont suffisamment proches, les rassembler.

7)-Si la dispersion des objets autour d'une classe est trop importante, éclater la classe en deux.

8) –Si (il y a eu des modifications) ou (un certain nombre T d'itérations n'a pas été atteint) alors aller en 3.

9)-affecter chaque objets à la classes dont le centre est le plus proche.

10)-retourner la partition obtenue.

Fig1.5 Algorithme ISODATA

A la fin des années 90, de nouveaux algorithmes de classification automatique sont apparus avec d'autres principes. Parmi ces algorithmes on peut citer notamment les techniques d'évaluation de modèles statistiques, les techniques de modélisation par réseaux de neurones[8], les techniques basées sur différents concepts tel que la théorie

des graphes[9], recherche stochastique, la logique floue[10] et les méthodes hybride avec les métaheuristiques (recuit simulé, recherche tabou, algorithmes génétique…..).

I.4 Domaine d'application

Les algorithmes de classification peuvent être appliqués dans beaucoup de domaines, sa terminologie diffère d'un domaine à un autre par exemple en :

- Science naturelle : la taxinomie de définit comme « l'art ou la science de la classification » ;
- Biologie : classification des végétaux et des animaux en faisant apparaître leurs différents caractères.
- Marketing : cibler des groupes de clients ayant un comportement semblable, et cela grâce à une grande base de données de clients contenant leurs propriétés, leurs habitudes et méthodes d'achat.
- Médecine : la nosologie est la classification des maladies.
- Reconnaissance des formes : reconnaissance de la parole, vision artificielle ….etc.
- Bibliothèque : répartition des livres.
- Sécurité : reconnaissance des visages ou des empreintes digitales pour sécuriser l'accès à des endroits surveillés.
- WWW : classification de documents, de pages Web pour faciliter leur accès à l'aide de mots clefs.
- Etude de tremblement de terre : recensement des épicentres pour identifier des zones dangereuses.
- Ville-planification : identification des groupes de maisons selon leur architecture, valeur et endroit géographique.

I.5 Conclusion

Nous venons d'évoquer dans ce chapitre les différentes méthode de classification automatique, vu la multitude de méthodes qui laisse l'utilisateur perplexe devant le choix de l'une d'elles pour une application donnée parce que il n'y a pas de règle générale permettant de sélectionner une méthode particulière de classification pour un problème donné .Chacune des méthodes a des avantages et des inconvénients. C'est la raison qui nous a poussé à faire une étude comparative entre quelques méthodes de classification automatique pour extraire l'avantage d'une méthode paraport aux autres.

Chapitre II

Algorithmes des méthodes de classification

II.1 introduction

Comme on l'a vu précédemment dans le premier chapitre, il existe plusieurs méthodes de classification, cette multiplicité de méthodes met souvent l'utilisateur dans l'embarras du chois de la méthode adéquate, et il est souvent très difficile de faire ce choix, puisque il n'existe presque pas de protocole universelle pour le faire.

C'est pour cette raison qu'on s'est proposer de faire une étude comparative de quelques méthodes de classification afin d'en tirer les avantages et les inconvénients de ces méthodes pour faciliter la tâche du choix de ces dernières.

Donc dans ce chapitre on va exposer les algorithmes et principes des méthodes de classification les plus classique et les plus utilisées.

II-2 Algorithme de classification hiérarchique ascendante

L'algorithme est du à Sokal [11], puis étudié par Lance et Williams [12], et Gordon [13]. Elle nous fournie pas une partition mais une hiérarchie de partition. La classification hiérarchique se représente par un dendrogramme ou arbre de classification.

- Principe

Soit H un ensemble de m individus tel que :

H=$\{I_1, I_2, \dots I_i, \dots \dots \dots, I_m\}$

Et pour chaque individu on affecte n attribues X_j ; j=1……..n

Le principe de l'algorithme consiste à créer, à chaque étape, une partition obtenue en agrégeant deux à deux les éléments les plus proches. On désignera alors par *élément* à la fois les individus ou objets à classer eux-mêmes et les regroupements d'individus générés par l'algorithme.

On commence tous d'abord par le calcule de la matrice des distances qui est une matrice (m x m) dont les éléments sont les distances euclidiennes entre tous les individus :

$$D(I_i, I_p) = \left(\left(X_{i,1} - X_{p,1}\right)^2 + \cdots + \left(X_{i,j} - X_{p,j}\right)^2 + \cdots + \left(X_{i,n} - X_{p,n}\right)^2\right)^{1/2} \qquad (2.1)$$

Où (i, p)= (1, 1)…….. (m, m) , j=1……..n

Une fois la matrice des distances est calculée on cherche la distance la plus petite et on détecte quelles sont les éléments qui sont séparés par cette distance, puis on les agrège pour former le premier groupe (en même temps il est considéré comme étant un individu). Comme ça on aura un ensemble de (m-1) individus .on construit la nouvelle matrice des distances de dimension (m-1) x (m-1) en prenant comme distance entre les individus et le groupe formé précédemment le **min** des distances entre l'élément en question et les deux éléments qui constituent ce groupe ,c'est à dire : si g est le groupe (classe) formé de a et b et c l'élément en question ,

$$dist(c, g) = min\{dist(a, c), dist(b, c)\}. \qquad (2.2)$$

Les autres éléments de la nouvelle matrice des distances restent ceux de la matrice précédente.

On a aussi la distance entre deux groupes (classes) A et B définie comme suite :

$$dist(A, B) = \underset{a\in A, b\in B}{min}(d(a, b)) \qquad (2.3)$$

On répète le processus d'agrégation jusqu'à n'avoir qu'une seule classe formée de tous les individus.

- déroulement de l'algorithme

Au début de l'algorithme chaque individu forme une classe

1- Calcul des distances euclidiennes entre tout les m individus :

$D(I_i, I_p)=0$

Pour $i = 1$ *à* m *faire*

Pour $p = i + 1$ *à* m *faire*

Pour $j = 1$ *à* n *faire*

$$D(I_i, I_p) = D(I_i, I_p) + (X_{ij} - X_{pj})^2$$

fin pour

$$D(I_i, I_p) = \sqrt{D(I_i, I_p)}$$

fin pour

fin pour

2- Construction de la matrice de distance $dist$ '

$$pour\ i = 1\ à\ m\ \ faire$$

$$Pour\ j = 1\ à\ m\ faire$$

$$dist(i,j) = \mathrm{D}(\mathrm{I_i}, \mathrm{I_p})$$

$$fin\ pour$$

$$fin\ pour$$

- On obtient à la fin une matrice distance $(m \times m)$ $dist$

$dist =$

0	.....	$\mathrm{D_{1.k}}$	.....	$\mathrm{D_{1.m}}$
.....	0	.....	.....	.....
$\mathrm{D_{k,1}}$	.....	0	.....	$\mathrm{D_{k,m}}$
.....	.....	.....	0	.....
$\mathrm{D_{m,1}}$	.....	$\mathrm{D_{m,k}}$	.....	0

t=0

Répète

t=t+1

1- *On cherche le minimum parmi les éléments de cette matrice* $dist$:

$$\min(t) = \text{minimum}\ \{dist(i,j)\}$$

2- On cherche les deux individus qui sont reliés par cette distance

3-On construit la première classe A composée des individus (i,j) :

A=$\{(i,j)\}$

-On remplace les deux vecteurs colonnes (i^{eme}, j^{eme}) et les deux vecteurs lignes (i^{eme}, j^{eme}) de la matrice de distance par un seul vecteur w tel que ce vecteur w a pour

composantes le maximum entre les composantes des vecteurs $i\ et\ j$. Ce vecteur w représente les composantes de la classe A dans la matrice des distances

4-on construit une nouvelle matrice de distance $(m-t)\times(m-t)$

Jusqu'à t=m-1

A la fin on obtient une classe qui contient le nombre d'individus à classer

Pour avoir un certain nombre k de classes on fixe le niveau d'agrégation on trace une ligne horizontale d'une manière à avoir k intersections avec le dendrogramme des classes

II-3 Algorithme de la méthode des K- means

Cet algorithme a été proposé par Macqueen en 1967 [14], il est l'un des algorithmes les plus utilisés en classification automatique.

- **Principe**

Soit un ensemble X de ***m*** individus caractérisés par ***n*** variables (attributs), à partitionner en K classes $\{c_1,\ldots\ldots c_k,\ldots\ldots\ldots c_K\}$.

Le principe de l'algorithme est de partitionner l'ensemble des points en un nombre de classes prédéterminé. Initialement un choix aléatoire est effectué sur K points qui vont constituer les centres provisoires des classes. Les autres points sont alors assignés à la classe dont le centre est le plus proche selon la distance euclidienne définie par la formule suivante :

$$D(X_i, V_k) = \|X_i - V_k\| = \sqrt{\sum_{j=1}^{n} (X_{i,j} - V_{k,j})^2} \qquad (2.4)$$

$$i= 1 \text{ à } m \ ; \ k=1 \text{ à } K$$

Avec : $D(X_i, V_k)$ représente la distance entre un individu X_i et le centre V_k.de la classe K

Une fois que tous les individus sont affectés a leurs classes qui convient, on recalcule les nouveaux centres de classes (il s'agit de calculer le centre de gravité des classes) et cela selon la formule suivante :

$V_{k,j} = (1 / N_k) \sum_{i=1}^{N_k}(X_{i,j})$ j= 1 à n (2.5)

Où N_k est le nombre d'observations de la classe C_k

Et on répète le même processus jusqu' à avoir une partition optimale, mais cette fois ci on n'attend pas l'affectation de tous les individus pour recalculer les nouveaux centres mais a chaque fois qu'un individu change de classe on le fait.

On a choisie comme critère d'arrêt, la stabilité des classes c'est-à-dire lorsque les centres des classes ne change pas durant deux d'itérations successives.

Pour le nombre de classe optimal K, on a choisis d'optimiser le critère VCR.il sera utilisé pour les trois méthodes de partitionnement

-Critère VCR

En classification non supervisée, on cherche généralement à partitionner l'espace en classes concentrées et isolées les unes des autres.

Dans cette optique, l'algorithme des K-means vise à minimiser la variance intra-classes, et à maximiser la variance inter-classe qui se traduit par la maximisation du critère VCR,

$$\text{VCR}=\frac{\sum_{k=1}^{K}\sum_{i=1}^{N_k}\sum_{j=1}^{n}(X_{i,j}-V_{k,j})^2}{\sum_{k=1}^{K}\sum_{j=1}^{n}(V_{k,j}-G_j)^2} \qquad (2.6)$$

où K est le nombre de classes, N_k le nombre d'éléments de la classe k , n, la dimension de l'espace de représentation et $X_{i,j}, V_{k,j}, et\ G_j$, respectivement les coordonnées des objets dans l'espace multidimensionnel, le centre de gravité de la classe k et le centre de gravité de tout le nuage des points sont calculés par la relation suivante.

$V_{k,j} = \frac{1}{N_k} \times \sum_{i=1}^{N_k} X_{i,j}$ (2 .7)

j= 1 à n

Une fois le calcule du VCR est fait s'il est bon (max) on arrête si non on recalcule les nouveaux centre et on réaffecte les individus a leurs classe qui convient, jusqu'à avoir le VRC max.

- Algorithme

1-détermination de nombre de classe optimal K_{opt}

Pour k=2à10 faire

$$VCR(k)= \frac{\sum_{k=1}^{K}\sum_{i=1}^{N_k}\sum_{j=1}^{n}(X_{i,j}-V_{k,j})^2}{\sum_{k=1}^{K}\sum_{j=1}^{n}(V_{k,j}-G_j)^2}$$

Fin pour

VCR_{max}=max (VCR(k))

K_{opt}=k

2) application de la méthode des k=means

–Tirer K_{opt} centres initiaux d'une manière aléatoire

- Affectation des observations aux classes C_k ;

Calculer les distances euclidienne D(Xi , V s) , ∀ s=1…….K

$$D(X_I, V_s) = \|X_i - V_s\| = \sqrt{\sum (X_{i,j} - V_{s,j})^2} \qquad j=1 \text{ à } n;\ i= 1 \text{ à } m$$

Affecter X_i aux classes C_s telque : $d(X_i, V_s) \leq d(X_I, V_k)$

- Calculer les nouveaux centres

$$V_{kj} = (1/N_k) \sum X_{ij}$$

- calculer les nouvelles distances euclidienne D (Xi, V s) par rapport aux nouveaux centres

Réaffecter les points par rapport aux nouveaux centres tel que: $d(X_i, V_s) \leq d(X_I, V_k)$ k=1…….K; s ≠ k.

- Si les centres ne changent pas au bout de trois itérations alors on affichage les classes.

FIN.

II-4 L'algorithme de la méthode des centres mobiles

L'algorithme peut être imputé principalement à Forgy (1965) [5] , bien que de nombreux travaux (parfois antérieurs : Thorndike, 1953), le plus souvent postérieurs (MacQueen, 1967; Ball and Hall, 1967) ont été menés parallèlement et indépendamment pour introduire des variantes ou des généralisations.

- Principe

Soit un ensemble X de ***m*** individus caractérisés par ***n*** variables (attributs), à partitionner en K classes $\{c_1,\ldots\ldots c_k,\ldots\ldots c_K\}$.

Le principe de l'algorithme est de partitionner l'ensemble des points en un ensemble de classes prédéterminé. Des points initiaux (V_k) sont choisis aléatoirement pour constituer les centres provisoires des classes. Les autres points sont alors assignés à la classe dont le centre est le plus proche selon la distance euclidienne.

Une fois que tous les individus sont affectés a leurs classes qui convient tel que :

$$D(X_i, V_s) \leq d(X_I, V_k) \quad k=1\ldots\ldots K;\ s \neq k$$

On recalcule les nouveaux centres de classes et cela selon la formule suivante :

$$V_{k,j} = (1 \ / \ N_k) \sum_{i=1}^{N_k} (X_{i,j}) \qquad (2.8)$$

j= 1 à n

Où N_k est le nombre d'observations de la classe C_k

Et on répète le même processus jusqu' à satisfaction d'un critère d'arrêt.

Et comme on le remarque la seule différence qui existe entre la méthode des K-means et celle des centre mobiles est la mise à jour des centres, dans la méthode des K-means on recalcule les centres à chaque fois qu'un individu change de classe mais dans les centres mobiles on le fait une fois que tous les individus sont réaffectés.

II-5 ISODATA

ISODATA est un algorithme qui permet au nombre de classes d'être ajusté automatiquement pendant le déroulement de l'algorithme, en fusionnant les classes les plus semblables et séparer celle qui sont dissemblable, mais ça se fait avec le réglage adéquat des paramètres de la méthode.

- **principe :**

1) le réglage des paramètres suivants :

ON : seuil du nombre d'éléments pour l'élimination d'un groupe.

OC : seuil de la distance pour l'union des groupes.

OS : seuil d'écart type pour la division d'un groupe.

K : nombre (maximum) de groupes.

I : nombre maximum des itérations permises.

2) L'algorithme ISODATA construit une partition initiale de K classes, en tirant au hasard K centres provisoires et affecte les individus à la classe dont le centre est le plus proche, en utilisant comme distance la distance euclidienne définie comme suite :

$$D(X_I, V_s) = \|X_i - V_S\| = \sqrt{\sum_{j=1}^{n}(X_{i,j} - V_{s,j})^2}\ ; \qquad (2.9)$$

I= 1 à m ; s= 1 à K

X_i:étant l'individu en question

V_s : le centre de la classe s

3) Ensuite, les classes sont recomposées en suivant un certain nombre de règles :

- On fait **L'éclatement** (dédoublement) d'une classe si cette classe a une dispersion maximale dépassant un seuil donné OS.
- $\sigma(s) = \sqrt{\frac{1}{N_s}\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{i,j} - V_{s,j})^2}$ (2.10)
- $s = 1$ à K;

- **Agglomération** de deux classes s′ et s″ si la distance entre leurs centres est $V_{s'}$ et $V_{s''}$ est inférieure à un seuil donné **OC**

- **Suppression** d'une classe 's' si son effectif est inférieur à un seuil donné ON

- algorithme :

1 - Tirer K centres initiaux d'une manière aléatoire

2- Affectation des observations aux classes C_k ;

Calculer les distances euclidienne D (Xi , V s) , ∀ s=1……K

$$D(X_I, V_s) = \|X_i - V_s\| = \sqrt{\sum (X_{i,j} - V_{s,j})^2}$$

3- Affecter X_i à la classes C s tell que: $D(X_i, V_s) \leq d(X_i, V_k)$ k=1…….K; s ≠ k.

4-éliminer les classes qui ont un nombre d'individu $Ns < O_N$ (alors les individus de ces classes deviennent orphelins)

5-affecter les individus orphelin aux classes dont le centre est le plus proche.

6-calcule des centres de gravité des classes :

7-réaffecter les individus par apport aux nouveaux centres.

8. le fusionnement des classes

Pour s=1 à K faire

Pour r=(s+1) à K faire

$$D(V_s, V_r) = \|X_s - V_r\| = \sqrt{\sum (X_{sj} - V_{rj})^2}$$

Fin pour

Fin pour

Si $D(V_s, V_r) < Oc$ fusionnez C (s) et c(r) (fusionnement de deux classes)

K=k-1

Si non ne rien faire.

9 -On trouve le composant maximum de chaque σ_s et on le note $\sigma^{(s)}_{max.}$

Pour i=1 à m faire

Pour s=1 à K faire

$$\sigma(s) = 0$$

Pour j=1 à n faire

$$\sigma(s) = \sigma(s) + \left(X_{i,j} - V_{s,j}\right)^2$$

$$\sigma(s) = \sqrt{\frac{1}{N_s}\sigma(s)}$$

$fin\ pour$

$fin\ \ pour$

$\sigma max = max\sigma(s)$

Si $\sigma max > OS$

Alors on calcul V_{s-} et V_{s+}

Si $\sigma^{(s)} j > O_S$; alors deviser la classe 's'en deux classes.

K=k+1

Si non ne rien faire

10-s'il y a eu des modifications ou I n'est pas atteint alors aller à 3.

11-affecter chaque objets à la classe dont le centre est le plus proche.

12-retourner la partition obtenue.

.**Fin**

Conclusion :

Dans ce chapitre on a essayé de détailler quelques algorithmes qui sont considérés comme étant des algorithmes très utilisés en ce moment en l'occurrence l'algorithme hiérarchique ascendant, la méthode des K-means, la méthode d'agrégation autour des centres mobiles et enfin l'algorithme ISODATA.

Chacune de ces méthodes a ses avantages et ses inconvénients .Par exemple la méthode ISODATA possède beaucoup de paramètres a spécifiés, chose qui est à la fois un avantage et un inconvénient ; c'est un avantage dans le sens ou ils nous permettent d'avoir une meilleure partition grâce aux conditions (paramètres) imposées par l'algorithme.

C'est un inconvénient dans le sens ou l'utilisateur doit être un expert dans le domaine de l'application de cette méthode pour savoir comment fixer ces différents paramètres, chose qui n'est pas toujours vérifiée.

Pour mieux comprendre et mieux voir les avantages et les inconvénients de ces méthodes, dans le prochain chapitre on va exposer les résultats obtenus avec les différents programmes et les tests sur quelques fichiers de données générés d'une façon artificielle, cela va nous permettre de bien comparer les différentes méthodes et extraire les avantages et les inconvénients.

Chapitre III

Tests et résultats

III-1 Introduction :

Les tests sur les différents algorithmes de la classification automatique étudiés en détail dans le chapitre précédent sont avérés importants, et c'est pour cette raison qu'on a fait une série de test sur les trois méthodes les plus utilisées en classification automatique

(K-means, isodata, et hiérarchique ascendante), et selon les différents résultats on tire les déférents avantages et inconvénients de chaque méthodes et leur limite en pratique. En exécutant les programmes sous MATLAB de chaque méthode sur trois fichiers de taille déférente (faible, moyenne, et grande).

III-2 Aperçue sur le logiciel MATLAB:

MATLAB est un logiciel de calcul scientifique dédie plus particulièrement à l'application numérique. À l'origine il a été conçu pour manipuler des données mathématique ce qui on fait un outil majeur de l'analyse de données de traitement d'image et simulation numérique…etc. Il possède son propre langage de programmation avec des nombreuses fonctions fournées.

III-3 Tests et résultat sur les méthodes de classification

III-3-1 premier fichier.

Le premier fichier représente un tableau des mesures sur des produits ou les

Attribues (1) est une pression, et l'attribue(2) est un volume.

-Application sur la méthode hiérarchique

On a utilisé la méthode de saut maximal (complète-Link)

Le teste de premier fichier par la méthode hiérarchique ascendante a donné le résultat suivants :

1) Les résultats obtenus par notre programme sous MATLAB sont :

   Au début on à 21 classes ou chaque individu forme une classe.

   Les crochets représentent les classes et entre chaque crochet on trouve les individus de la même classe.

Classes initiales : {[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20] [21]}

1ere agrégation: {[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12 19] [13] [14] [15] [16] [17] [18] [20] [21]}

2eme agrégation::{[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12 19] [13] [14] [15] [16] [17 18] [20] [21]}

3eme agrégation:{[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12 19] [13] [14] [15] [16] [17 18] [20 21]}

4eme agrégation: :{[1] [2] [3] [4 5] [6] [7] [8] [9] [10] [11] [12 19] [13] [14] [15] [16] [17 18] [20 21]}

5eme agrégation:{[1] [2] [3] [4 5] [6] [7] [8] [9] [10] [11] [12 19] [13 16] [14] [15] [17 18] [20 21]}

6eme agrégation:{[1] [2] [3] [4 5] [6] [7] [8 10] [9] [11] [12 19] [13 16] [14] [15] [17 18] [20 21]}

7eme agrégation :{[1] [2] [3 17 18] [4 5] [6] [7] [8 10] [9] [11] [12 19] [13 16] [14] [15] [20 21]}

8eme agrégation :{[1] [2] [3 17 18] [4 5] [6 12 19] [7] [8 10] [9] [11] [13 16] [14] [15] [20 21]}

9eme agrégation: {[1] [2] [3 17 18] [4 5] [6 12 19] [7] [8 10] [9 11] [13 16] [14] [15] [20 21]}

10eme agrégation :{[1] [2] [3 17 18] [4 5] [6 12 19] [7] [8 10 20 21] [9 11] [13 16] [14] [15]}

11eme agrégation :{[1] [2 3 17 18] [4 5] [6 12 19] [7] [8 10 20 21] [9 11] [13 16] [14] [15]}

12eme agrégation: {[1] [2 3 17 18] [4 5] [6 12 19] [7] [8 10 20 21] [9 11] [13 16 15] [14]}

13emeagrégation: {[1] [2 3 17 18] [4 5] [6 12 19] [7 14] [8 10 20 21] [9 11] [13 16 15]}

14emeagrégation: {[1] [2 3 17 18] [4 5] [6 12 19 9 11] [7 14] [8 10 20 21] [13 16 15]}

15emeagrégation: {[1] [2 3 17 18] [4 5 8 10 20 21] [6 12 19 9 11] [7 14] [13 16 15]}

16emeagrégation: {[1] [2 3 17 18 7 14] [4 5 8 10 20 21] [6 12 19 9 11] [13 16 15]}

17emeagrégation : {[1] [2 3 17 18 7 14] [4 5 8 10 20 21] [6 12 19 9 11 13 16 15]}

18emeagrégation: {[1] [2 3 17 18 7 14 4 5 8 10 20 21] [6 12 19 9 11 13 16 15]}

19emeagrégation:{[1] [2 3 17 18 7 14 4 5 8 10 20 21 6 12 19 9 11 13 16 15]}

20emeagrégation: {[1 2 3 17 18 7 14 4 5 8 10 20 21 6 12 19 9 11 13 16 15]}

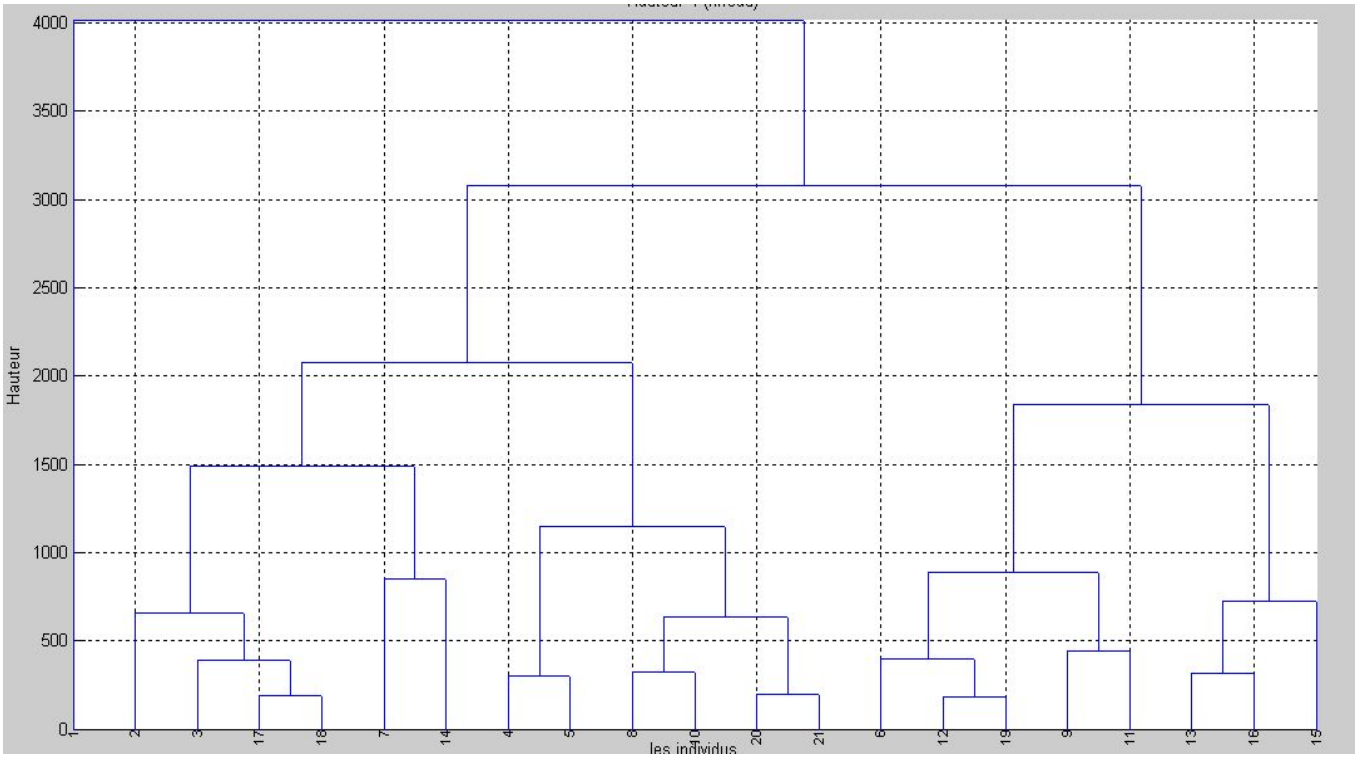


fig.3.1 Dendrogramme des classes sur le premier fichier est donné dans la ou l'axe X représente les classes et les Y le niveau d'agrégation des classes.

-Le programme de la méthode hiérarchique pour le premier fichier a duré : 0.53 secondes.

-application sur ISODATA.

-Signification des paramètres de la fonction.

ON= seuil du nombre d'éléments pour l'élimination d'un groupe.

OC= seuil de la distance pour l'union des groupes.

OS= seuil d'écart type pour la division d'un groupe.

k= nombre (maximum) de groupes.

L=3 : nombre maximum des groupes qui peuvent être mélangés dans une itération simple.

I=10: nombre maximum des itérations permises.

-Paramétrage de la fonction : ON=3;OC=1000;OS=20;k=7;L=20;I=30;

La méthode ISODATA a donné comme résultat sur le premier fichier :

-Les cordonnées des Centres estimés par ISODATA sur le premier fichier :

C1= 1.0e+003 (2.4020, 1.1386)

C2=1.0e+003 (0.8763, 0.2774)

C3=1.0e+003 (2.0209, 2.72)

C4= 1.0e+003 (1.6454, 1.1985)

- Les classes obtenues par la méthode ISODATA sur le premier fichier sont :

Classe1= {8 10 13 15 16}

Classe2= { 1 9 11}

Classe3= { 2 3 4 5 7 14 17 18}

Classe4= { 6 12 19 20 21 }

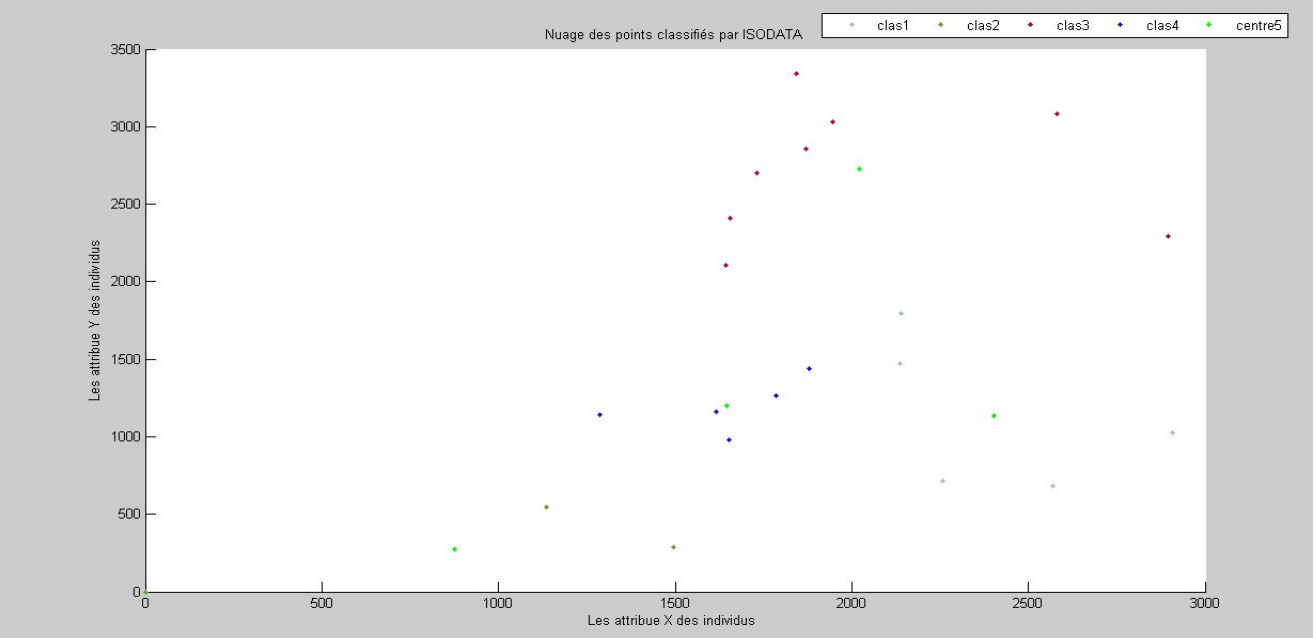


Fig. 3.2 : Nuage des points obtenu par ISODATA sur le premier fichier.

-Programme ISODATA pour le premier fichier a duré : 0.33 secondes.

- application sur k-means.

-Le nombre de centre optimal est : k= 4.

-Les cordonnées des centres estimés par la méthode k-means pour le premier fichier sont :

C1= 1.0e+003 × (0.8763 ,0.2774)

C2=1.0e+003 × (2.0114, 0.9967)

C3=1.0e+003 × (1.9503, 1.7041)

C4=1.0e+003× (2.0748, 2.8155)

-Les classes estimées par k-means sur le premier fichier

Classe1= {1 9 11}

Classe2= {6 12 13 15 16 19 20}

Classe3= {4 8 10 21}

Classe4= {2 3 5 7 14 17 18}

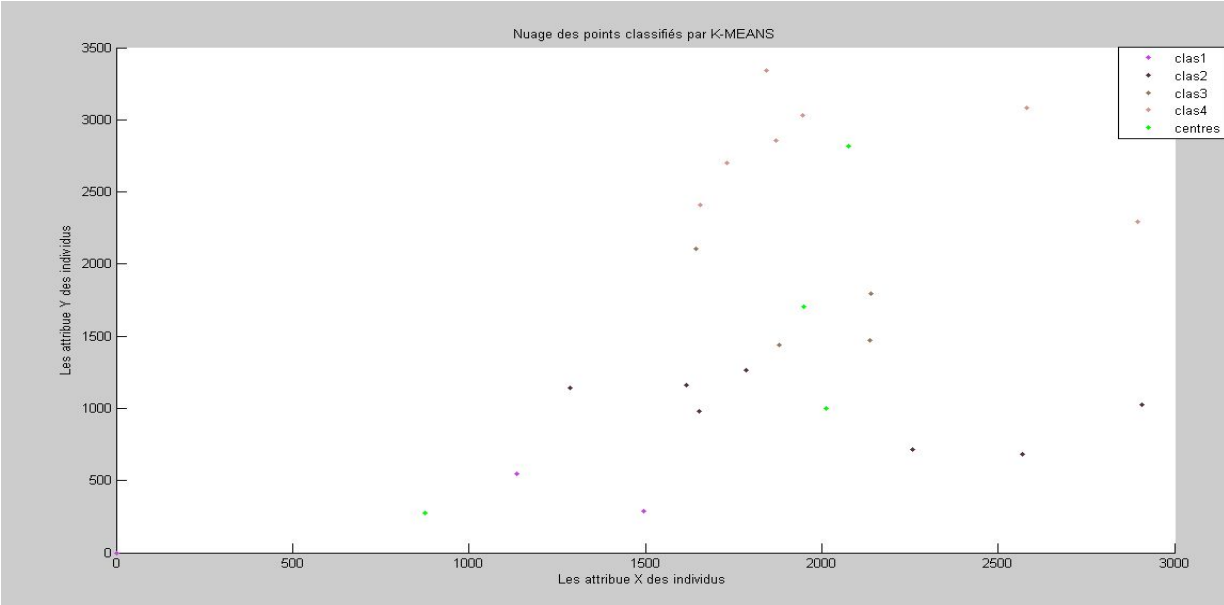


fig3.3 **:** Nuage du point obtenu K-means sur le premier fichier**.**

-Le programme k-means sur le premier fichier a duré 0.19 secondes.

III-3-2 Le deuxième fichier

Représente des observations faites sur 265 individus et chaque individus possède 2 attribues.

-Application sur hiérarchique ascendante

1) Les classes : il existe 264 agrégations des classes, et à cause de volume important de résultat on donne seulement le nombre de classes pour chaque étape d'agrégation ou fusionnement.

   Le programme fait une agrégation pour chaque étape ou il fusionne deux classes pour construire une nouvelle classe jusqu'à avoir une seule classe à la 264^{eme} agrégation.

   Au début on a 265 classes ou chaque individu représente une classe.

-Apres la première agrégation on obtient 264 classes.

-Apres la deuxième agrégation on obtient 263 classes.

-Apres la troisième agrégation on obtient 262 classes.

-Le programme continue les agrégations jusqu'à la 264ème agrégation.

Dans cette dernière étape le programme nous donne une seule classe de 265 individus.

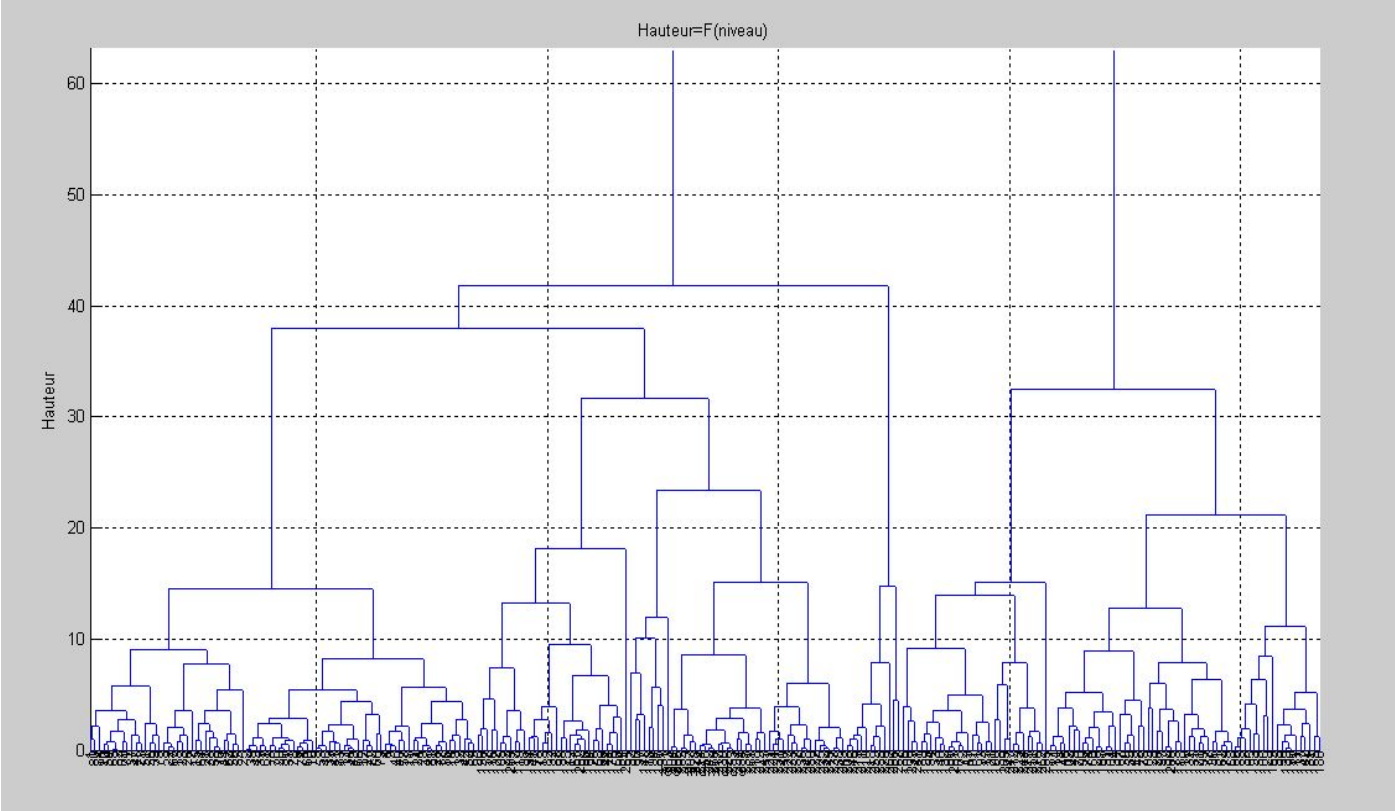


fig.3.4 : Dendrogramme des classes fusionnées de deuxième fichier.

-Programme de la méthode hiérarchique ascendant a duré 1.98 secondes.

-Application sur ISODATA.

-La signification des paramètres sont cités dans le premier teste sur la méthode pour le premier fichier.

-Paramétrage de la fonction : ON=10; OC=11; OS=0.8; k=7; L=8; I=30

- Les cordonnées des centres estimées par la méthode :

C1= (9.6197, 45.9107)

C2= (4.7009, 6.3864)

C3= (23.9526, 21.2436)

C4= (10.6958, 40.7780)

- Les classes obtenues par la méthode ISODATA sur le deuxième fichier:

Classe1= {2 3 4 5 6 7 8 9 12 13 15 16 17 18 20 21 22 23 25 26 27 30 31 33 34 35 36 37 38 39 40 41 43 45 46 49 50 53 55 57 58 60 61 62 64 66 67 68 70 72 74 76 78 79 83}

Classe2={ 85 86 87 88 89 90 91 93 94 95 99 100 101 102 103 104 105 106 107 108 109 110 111 113 114 115 116 117 119 120 121 122 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 165 166 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 185 187 188 189 190 192 193 195 197 198 199 200 201 203 209 239 241 247 }

Classe3= {96 97 98 112 123 147 164 167 184 196 202 204 205 206 207 208 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 240 242 243 244 245 246 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 265}

Classe4= {1 10 11 14 19 24 28 29 32 42 44 47 48 51 52 54 56 59 63 65 69 71 73 75 77 80 81 82 84 92 118 186 191 194 264 26}

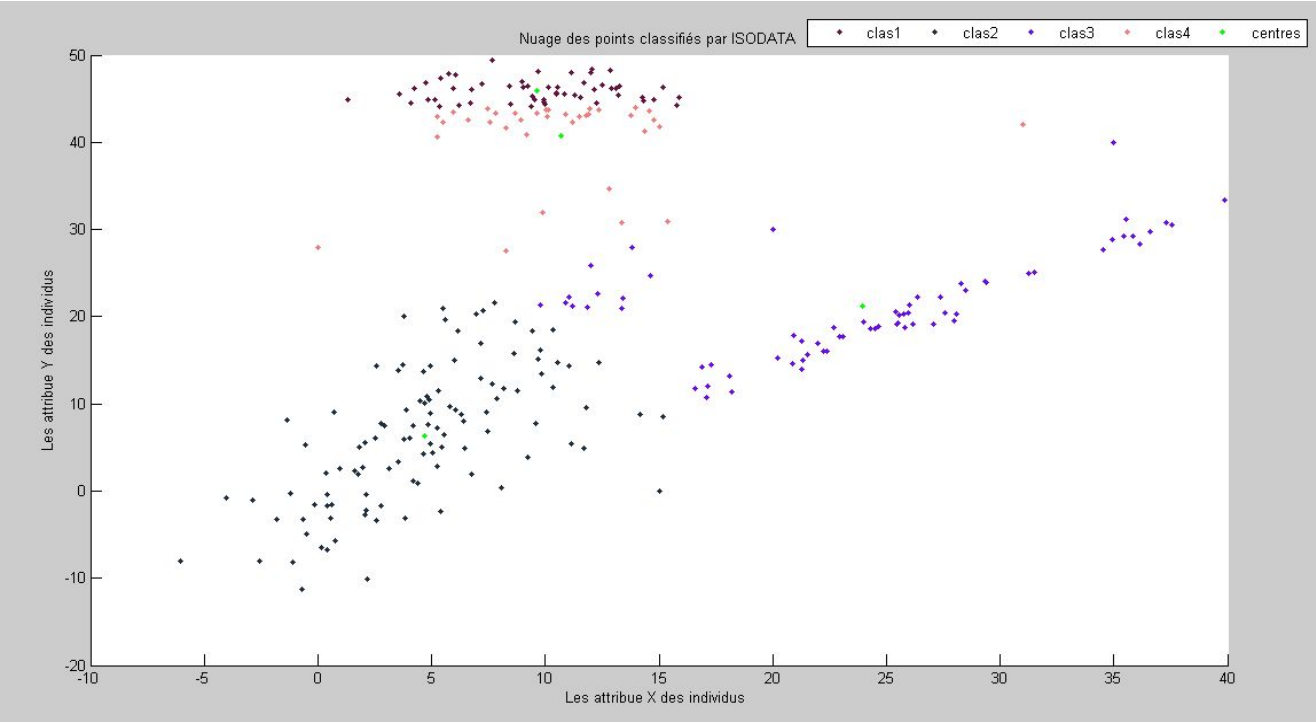


fig.3.5 Nuage du point obtenu par ISODATA sur le deuxième fichier.

Programme ISODATA pour deuxième fichier a duré : 1.03 secondes.

-Application sur k-means.

-Nombre des centres optimal est k=4.

-Les cordonnées des centres estimés par la méthode sont :

C1= (2.7901 ,1.3755)

C2= (9.3670 ,15.7984)

C3= (27.3013 ,22.6543).

C4= (9.8789 ,44.4215).

Les classes estimées par la méthode k-means sur troisième fichier sont:

classe1= {87 88 90 91 93 95 99 102 103 104 106 107 109 110 111 114
120 124 125 128 129 130 131 133 134 135 136 137 138 139 142 145
146 148 149 151 152 155 159 160 162 165 168 169 170 171 173 174

175 177 179 180 181 183 185 188 189 190 195 197 199 200 239 241 265}

classe2= {85 86 89 92 94 96 97 98 100 101 105 108 112 113 115 116 117 119 121 122 123 126 127 132 140 141 143 144 147 150 153 154 156 157 158 161 163 164 166 167 172 176 178 182 184 187 192 193 196 198 201 202 203 209 212 215 216 223 244 247 251 252 264}

classe3= { 118 204 205 206 207 208 210 211 213 214 217 218 219 220 221 222 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 240 242 243 245 246 248 249 250 253 254 255 256 257 258 259 260 261 262 263 266 252}

classe4={ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 186 191 194}

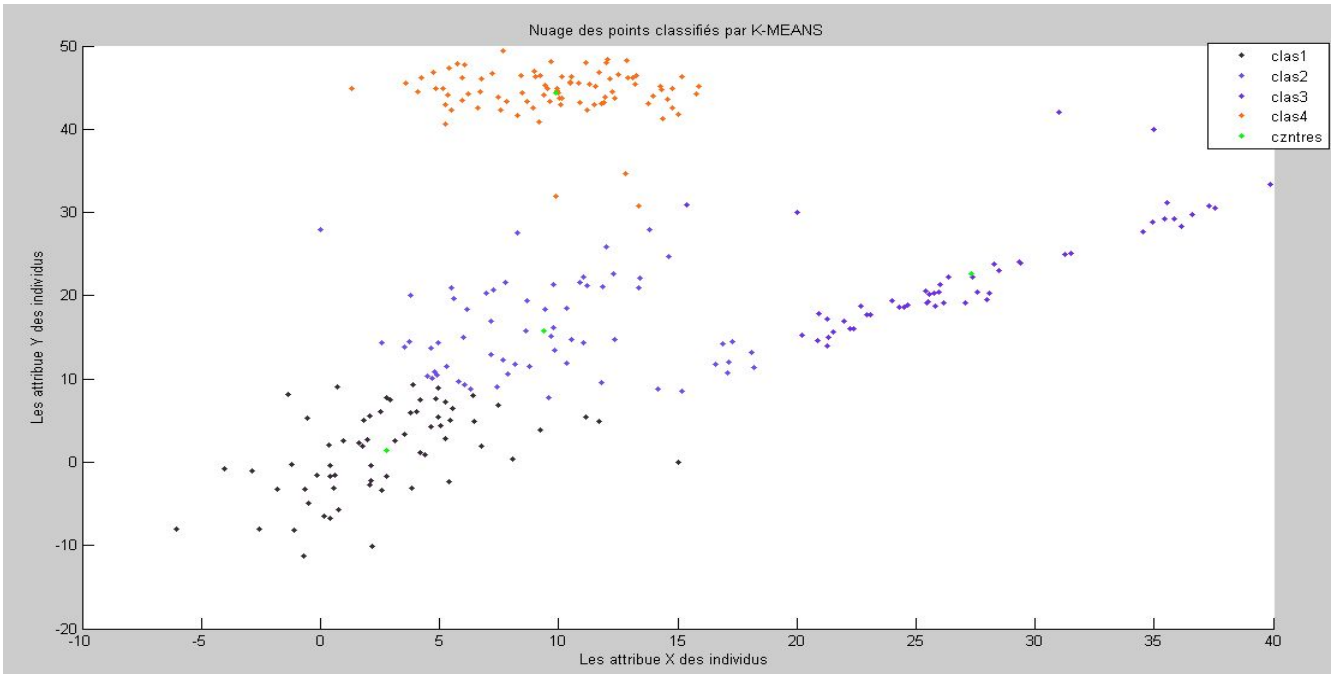


fig3.6 : Nuage du point obtenu par K-means sur le deuxième fichier

Le programme k-means sur le deuxième fichier à duré:0.4063s

III-3 -3 Troisième fichier.

Contient un tableau de 4000 individus et chaque individu possède 2 attribues.

-Application sur méthode hiérarchique: pas de résultat pendant 12 minutes après MATLAB nous a envoyé message d'erreur (mémoire insuffisante).

-application sur méthode ISODATA.

-Paramétrage de la fonction

ON=200; OC=11; OS=0.8; k=7; L=3; I=40;

-La méthode a estimée les résultats suivants :

Nombre de classe estimés dans le troisième fichier par la méthode ISODATA est : 2.

Les cordonnées des centres estimées par ISODATA sur le troisième fichier sont :

C1= (0.0001, 0.1242)

C2= (-0.0001, -0.123)

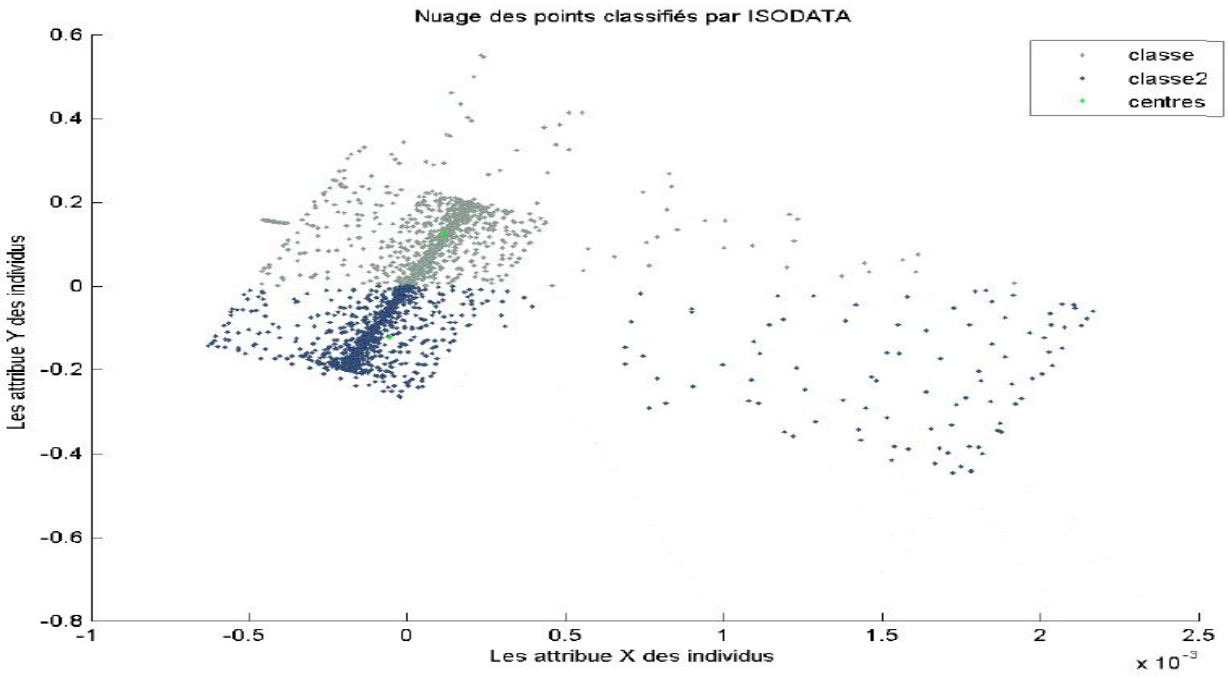


fig3.7Nuage des points obtenu par ISODATA sur le troisième fichier

Programme ISODATA sur le troisième fichier a duré : 9.77 secondes.

-Application sur K-means.

Nombre de centres à trouver est 3.

Nombre de classes estimées par la méthode k-means sur le troisième fichier est est : 3.

Les cordonnées des centres estimées par la méthode k-means sue le troisième fichier est :

C1= (-0.0001, -0.1586)

C2= (0.0000, -0.0007)

C3= (0.0001, 0.1578)

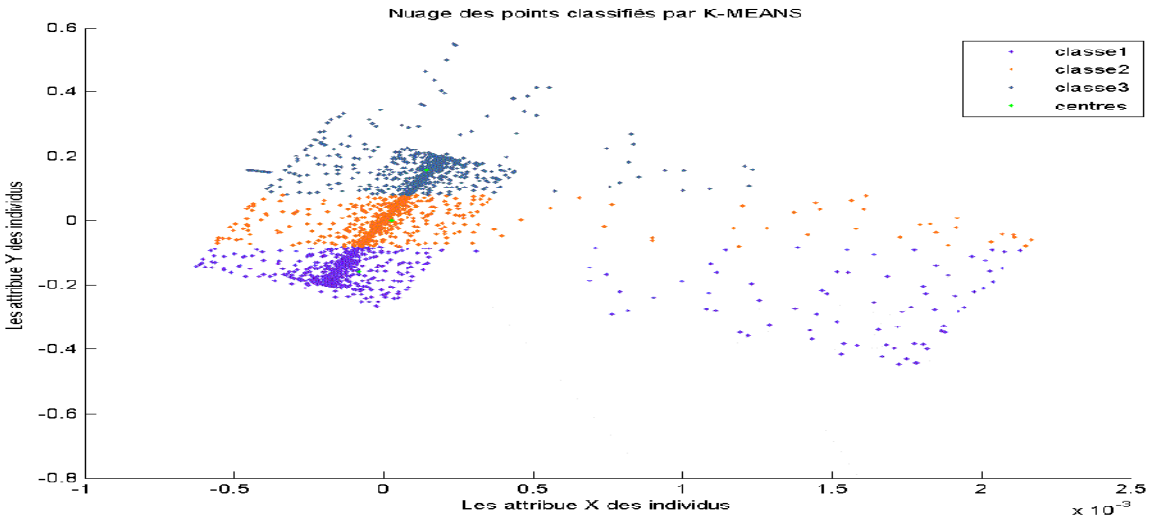


fig.3.8 Nuage du point obtenu par K-means sur le troisième fichier

Programme k-means sur le troisième fichier a duré : 5.91 secondes

III-4 Evaluation du résultat

-hiérarchique ascendante :

Cette méthode a comme avantage :

-Elle nous donne le nombre de classe maximal qu'on peut trouver dans fichier de donnée

-elle nous permet la visualisation des agrégations de toute la paire de classe d'une manière successive

Et a comme inconvénient :

-Dans le traitement du fichier de grande taille ou elle prend un temps suffisamment grand, et sa convergence devient trop lente et

- L'exécution de son algorithme sur les fichiers de grande taille a besoin d'un grand espace mémoire et son dendrogramme devient difficile à interpréter.

- deux classes agrégées ne seront jamais disponible.

- méthode ISODATA :

-Cette méthode a comme avantage : elle est plus autonome que les autres méthodes de partitionnements puisque elle n'a pas besoin de calculer les nombre de classe optimal, car ils sont optimisés d'une manière automatique par son algorithme.

- Elle nous donne les classes et leur centres au contraire à la méthode hiérarchique.

-Sont algorithme utilise tout le fichier de donnée d'une manière itérative pour trouver la classification optimale.

-Sont algorithme réajuste les centre et leur nombre au même temps.

Et ses inconvénients :

-Trop de paramètres internes et la difficulté de leurs paramétrage car un faux paramètre peut diminuer la qualité des résultats et augmente le temps d'exécution de programme (convergence lente).

-Sa vitesse de convergence est légèrement moins rapide que k-means sur le fichier de taille importante à cause du paramétrage car ce dernier dépend de la caractéristique du fichier.

- méthode k-means.

Cette méthode a comme avantage :

-Elle nous donne les classes et leurs centres au contraire à la méthode hiérarchique.

-Peu de paramètres internes à donner au programme d'exécution.

-son inconvénient est son indépendance au nombre de centres initial.

III-5 Conclusion :

Apres les tests effectués sur les différentes méthodes par les trois fichiers de données de taille différentes on a constaté que les méthodes de partitionnements nous donnent un avantage sur les résultats par rapport aux méthodes hiérarchiques car cette dernière elle est limitée sur des fichiers de taille faible.

ISODATA est l'une des meilleures méthodes de classification automatique des données mais toujours après choix difficile de paramètre internes de cette méthode.

Conclusion générale

Conclusion générale

Le domaine de la classification automatique de données est vraiment très vaste, il trouve son application dans de nombreuse disciplines, il en a Profité du développement de l'outil informatique pour être un peu plus efficace.

Dans ce travaille on a exposé tout d'abord d'une manière générale les méthodes de classification automatique de données qui existent en ce moment, puis on a détaillé quelques méthodes qu'on a jugé les plus utilisées, et après on a testé ces méthodes sur quelques fichiers de données artificielle dont on connais déjà toutes les informations nécessaires, ce qui nous a permis de juger l'efficacité de la méthode d'une manière plus objective ..

Après les différents testes nous avons remarqué que la méthode hiérarchique ascendante a comme avantage de nous donner une description claire du fichier grâce aux dendrogramme qui nous permet de voir d'une manière progressive la formation des classes, cependant il a comme inconvénient le temps du déroulement du programme qui est un peu long, chose qui n'arrange souvent pas l'utilisateur, et notons aussi que lorsque le fichier de données est important de point de vue nombre d'individus, le dendrogramme devient presque illisible. Par contre la méthode des K-Means est beaucoup plus rapide, chose qui nous permet de traiter des fichiers de grand taille, mais a pour inconvénient la spécification du nombre de classes K, qui est souvent une tache difficile surtout lorsque on ne possède pas des informations à priori sur le fichier, mais quand même on a essayé de remédier à cela en introduisant le critère de validité dans le programme, chose qui nous a permit de déterminer K d'une manière automatique et d'avoir la partition optimale en même temps. Enfin la méthode ISODATA est plus sophistiqué elle nous permet d'avoir une meilleur partition avec un temps d'exécution acceptable, mais le problème qui se pose c'est la spécification des différents paramètres

Enfin ce travail nous a ouvert une fenêtre dans le domaine de la classification automatique de données, et nous a permis d'approfondir nos connaissances sur le langage MATLAB.

BIBLIOGRAPHIE

1] Y, Collette, P.Siarry «Optimisation multiobjective » édition Eyrolles. N⁰ 11168.2002

[2]El-Djillali Talbi « Sélection et Réglage de Paramètres pour L'optimisation de Logiciels D'ordonnancement Industriel » thèse de doctorat à l'Institut National Polytechnique de T

[3.] S. Lin, B.W. Kernighan, «An efficient heuristic for the traveling-salesman problem». Operations Research 21 : 498-516, 1973.oulouse.2004.

[5] Forgy, E. W. Cluster Analysis of Multivariate Data : E_ciency versus Interpretability Models. Biometrics, 61(3):768_769, 1965.

[6]. M. Dorigo, G. Di Caro. «The Ant Colony Optimization meta-heuristic». In D. Corne, M. Dorigo, and F. Glover, editors, New Ideas in Optimization, McGraw Hill, London, UK, 1999.

[7]. J. Kennedy . «Small Worlds and Mega-Minds : Effects of Neighborhood Topology on Particle Swarm» Performance. In IEEE Congress on Evolutionary Computation, volume III, pages 1932–1938.1999

[8]. F.van den Bergh, « An Analysis of Particle Swarm Optimizers. » PhD thesis, Department of Computer Science, University of Pretoria.2002.

[9]. M. Clerc, J. Kennedy, «The Particle Swarm : Explosion, Stability, and Convergence in a Multi-Dimensional Complex Space. » In Proceedings of the IEEE Transactions on volutionary Computation, volume VI, pages 58–73.2002.

[10]. J. H. Holland,«Adaptation in Natural and Artificial Systems »The University of Michigan Press : Ann Arbor, 1975.

[11] [SOK 63] SOKAL, R.R., SNEATH P.H.A. Principles of Numerical Taxonomy, Freeman and co., San

Francisco, 1963.

[12] [LAN 67] LANCE, G.N., WILLIAMS, W.T., A General Theory of Classification Sorting Strategies, Computer

Journal, vol.9, 373-380, 1967.

[13] [GOR 87] GORDON, A.D., A Review of Hierarchical Classification, J.R Statistics Soc., A, vol. 150, Part2,

119-137, 1987.

[14] MCQUEEN J. (1967) : Some methods for classification and analysis of multivariate

observations. Procs of 5th Berkeley Symposium on Math statistics and probability, pp 281-

297.