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

République Algérienne Démocratique et Populaire
Université Abou Bakr Belkaid– Tlemcen
Faculté de Technologie
Département d'Informatique

Mémoire de fin d'études

pour l'obtention du diplôme de Licence en Informatique

Thème

Classification non supervisée : Application de k-means

Réalisé par :

- Melle Oumiloud Horiya
- M^elle Mokeddem Asma

Encadré par :

- M^me CHAOUCHE L.

Présenté le 08 Juin 2014 devant la commission d'examination composée de :

- *M^r BENAISSA.M (Examinateur)*
- *M^r BRIKCI.A (Examinateur)*

Année universitaire : 2013-2014

REMERCÎMENT

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

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

A ce titre, nous remercions vivement Notre encadreur Mme Chaouche.L pour ses conseils et son suivi durant la réalisation de notre projet.

.

Aussi, nous tenons à exprimer notre reconnaissance aux membres du jury :M^r Benaissa.M et M^r Brikci.A

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

Dédicace

A qui m'as fait élever par une bonne éducation

A qui m'a allumé le chemin de savoir depuis mon enfance jusqu'à ma soutenance pour que je puisse atteindre mon objectif

A qui m'a procuré le confort, la paix et la béatitude éternelle

A qui a tout fait pour moi

Mes très chers parents que dieu les gardes

A qui j'ai passé avec eux les meilleurs souvenirs de ma vie
Mon cher frére Abdeljalil
Et mes jolies sœurs Soumia et Manel

A qui m'a partagé les bons et les mauvais moments pour réussir ce travail
A toi ma chère Asma et sa famille

A mes chers amis

Hoda, Khadidja, Amina, Zineb, Hayat.

Houria

Dédicace

A qui m'as fait élever par une bonne éducation

A qui m'a allumé le chemin de savoir depuis mon enfance jusqu'à ma soutenance pour que je puisse atteindre mon objectif

A qui m'a procuré le confort, la paix et la béatitude éternelle

A qui a tout fait pour moi

Mes très chers parents que dieu les gardes

A qui m'a donnée le courage l'amour et l'aide

Ma sœur Amina, mon frère Omar

A mes oncles, tantes et mes cousins mes cousines

je vous remercie pour votre Soutien

A qui j'ai passé avec eux les meilleurs souvenirs de ma vie

Mon chère Mohammed

mes amis Amel, Houria , Zeineb , Rahma ,Samia

A toute la promotion

ASMA

Table des matières :

Table des figures :

Résumé

La classification c'est construire une collection d'objets Similaires au sein d'un même groupe et dissimilaires quand ils appartiennent à des groupes différents. Les algorithmes de classification non supervisées sont souvent utilisés pour étudier des données pour lesquelles peu d'information sont disponible. Il existe une très large famille de méthodes dédiées à la classification non supervisée dont le plus simple est l'algorithme de k-means. Notre objectif de mémoire est d'appliquer une version de k-means sur un ensemble de données plus précisément un ensemble de points crées de façon aléatoire dans un espace à deux dimensions.

Abstract

Clustering is to construct a collection of Similar objects within the same group and dissimilar when they belong to different groups. Algorithms unsupervised classification are often used to examine data for which little information is available. There is a very large family of methods dedicated to unsupervised classification which the simplest is the k-means algorithm. Our memory objective is to apply a version of k-means on a data set and more exactly a set of points created randomly in a two-dimensional space.

ملخص

التجميع هو تكوين مجموعة من الأشياء متشابهة ضمن نفس المجموعة ومتباينة عندما تنتمي الى مجموعات مختلفة. وغالبا ما تستخدم خوارزميات التجميع الغير خاضعة للرقابة لفحص البيانات التي تتوفر على القليل من المعلومات المتاحة. هناك عائلة كبيرة جدا من الطرق المخصصة للتجميع الغير خاضع للرقابة أبسطها هو خوارزمية ذات المراكز المتنقلة. هدفنا في هذه الاطروحة هو تطبيق إصدار من خوارزمية ذات المراكز المتنقلة على مجموعة من البيانات و تحديدا مجموعة من العناصر التي تم إنشاؤها بشكل عشوائي في الفضاء ثنائي الأبعاد.

Chapitre I : la classification

I-1 Introduction :

La notion de classification est essentielle en science, elle permet aux scientifiques de mettre de l'ordre dans les connaissances qu'ils ont sur le monde .Ainsi, depuis longtemps des chercheurs et des scientifiques ont essayé de classer des espèces animales .De nombreuses classifications ont été crées .Face à ces classifications, les scientifiques sont souvent incapable de désigner la meilleure d'entre elles. Car chacune présente un intérêt par rapport aux autres et à la tâche considérée.

L'importance de la classification dans les sciences se reflète dans la grande variété des domaines ou tant leur nature que leur construction ont fait l'objet des recherches.

Dans le cadre des problèmes de classification, on dispose d'un ensemble de données qui reprend une collection d'individus (objet) non étiqueté. Les classes sont encore inexistantes. L'objectif alors d'obtenir des classes d'objets homogènes en favorisant l'hétérogénéité entre ses différentes classes [01].

Dans ce chapitre nous présenterons un panorama des méthodes de classification les plus connues et qui font. Référence à l'existence de groupes ou classes de données, elles se divisent en deux groupes:

• Les méthodes de classification automatique (aussi appelées méthodes de *clustering):* Méthodes basées sur la notion d'apprentissage non supervisé, laquelle consiste à regrouper des objets appartenant à un ensemble T en classes restreintes de telle sorte que les objets d'une même classe soient les moins dispersés possibles.

• Les méthodes d'affectation (aussi appelées «classificateurs») basées sur la notion d'apprentissage supervisé : méthodes utilisant un ensemble d'exemples où les classes d'appartenance sont connues au préalable. À partir de cet ensemble, des normes (ou règles) d'affectation seront définies.

De même, certains problèmes de classification nécessitent de combiner les deux types d'apprentissages (supervisée et non supervisée) appelé la méthode semi supervisé.

I-2 Domaines d'application de la classification :

La classification a un rôle à jouer dans toutes les sciences et techniques qui font appel à la statistique multidimensionnelle. Citons tout d'abord les sciences biologiques : botanique, zoologie, écologie,...Ces sciences utilisent également le terme de "taxinomie" pour désigner l'art de la classification. De même les sciences de la terre et des eaux : géologie, pédologie, géographie, étude des pollutions, font grand usage de classifications.

La classification est fort utile également dans les sciences de l'homme : psychologie, sociologie, linguistique, archéologie, histoire, etc. ... et dans les techniques dérivées comme les enquêtes d'opinion, le marketing, etc. ... Ces dernières emploient parfois les mots de "typologie" et "segmentation" pour désigner la classification, ou l'une de ses innombrables variantes. Citons encore la médecine, l'économie, l'agronomie, et nous en oublions certainement ! Dans toutes ces disciplines la classification peut être employée comme une fin en soi ; mais elle l'est souvent, à juste titre, comme une méthode complémentaire à d'autres méthodes statistiques. Elle peut, en effet, aider efficacement à l'interprétation des graphiques d'analyse factorielle, ou bien déterminer des groupes d'objets homogènes, préalablement à une régression linéaire multiple [02].

I-3 critères pour une bonne classification :

L'objectif principal des techniques de classification est de trouver une partition ou les objets d'une classe devraient être semblables (entre eux), les objets de différents classes devraient être différents [01], une bonne classification devrait accomplir différents critère :

I-3-1 Validité :

Elle peut se définir par :

- ***Chaque classe d'une partition doit être homogène*** *:*

Les objets qui appartiennent à la même classe doivent être semblables.

- ***Les classes doivent être isolées entre eux :***

Les objets de différentes classes doivent être différents.

- ***La classification doit s'adapter aux données :***

La classification doit pouvoir expliquer la variation des données.

I-3-2 Interprétabilité :

Les classes doivent avoir une interprétation substentive c'est-à-dire qu'il est possible de donner des noms aux classes ,dans le meilleur des cas les noms doivent correspondre aux types déduits d'une certaine théorie.

I-3-3 Stabilité :

Les classes doivent être stable ca veut dire les petites modification dans les données et dans les méthodes ne doivent pas changer les résultats.

I-3-4 D'autre critére :

Parfois la taille et le nombre de classes sont employés en tant que critéres additionnels :le nombre de classee doit étre aussi petit que possible , et la taille des classes ne doit pas étre trop petite [01].

I-4 Classification :

I-4-1 Classification supervisée :

C'est l'ensemble des techniques qui visent à deviner l'appartenance d'un individu à une classe en s'aidant uniquement des valeurs qu'il prend. Elle construire un modèle représentatif d'un certain nombre de données organisées en classes (ensemble) que l'on appelle généralement le corpus d'apprentissage - puis d'utiliser ce modèle afin de classer de nouvelles données, c'est à dire de prédire leur classe au vu de leurs caractéristiques (appelées paramètres ou features). La construction du modèle relève de l'apprentissage automatique supervisé, l'ensemble des exemples constituant le corpus d'apprentissage étant annotés, c'est à dire qu'ils portent le label de leur classe donné a priori.

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 [03] :

- **Méthode des k plus proches voisins**.
- **Réseau de neurones**.
- **Arbre de décision**.
- **Classification naïve bayesienne**.

La méthode directe de la classification supervisée k plus proches voisins.

1-4-1-1 K plus proches voisins (k-ppv) :

La méthode des plus proches voisins (noté parfois k-PPV ou k-NN pour (k-Nearest-Neighbor) consiste à déterminer pour chaque nouvel individu que l'on veut classer, la liste des plus proches voisins parmi les individus déjà classés. L'individu est affecté à la classe qui contient le plus d'individus parmi ces plus proches voisins. Cette méthode nécessite de choisir une distance, la plus classique est la distance euclidienne et le nombre de voisins à prendre en compte.

Cette méthode supervisée est souvent performante, cependant, le temps de prédiction est très long, car il nécessite le calcul de la distance avec tous les exemples, mais il existe des heuristiques pour réduire le nombre d'exemples à prendre en compte [04].

1-4-1-2 Affectation par la méthode bayésienne :

Un classificateur probabiliste linéaire simple basée sur le théorème de Bayes qui suppose que les descripteurs (attributs) qui décrivent les objets de l'ensemble d'apprentissage sont indépendants [02].

L'approche bayésienne a pour but de minimiser la probabilité d'erreur de classification, C'est-à-dire la probabilité jointe qu'une observation x soit en provenance d'une classe Ci et soit classée dans une autre [05].

1-4-1-3 Analyse discriminante :

Les méthodes d'analyse discriminante ont été largement étudiées; la littérature à ce sujet est très abondante.

Le but de ces méthodes est de produire des décisions concernant l'appartenance ou non d'un objet à une classe en utilisant des fonctions discriminantes appelées également *fonctions de décisions.* Suivant les formes des classes, on peut trouver différents types de discrimination: discrimination linéaire et discrimination quadratique [05].

1-4-1-4 Les réseaux de neurones :

Les réseaux de neurones sont à l'origine d'une tentative de modélisation mathématique du cerveau humain.

Le principe général des méthodes utilisant les réseaux de neurones consiste à modifier (ou ajuster) les paramètres comme, par exemple, le s poids et les seuils par des algorithmes itératifs afin d'obtenir des réponses correctes [05].

1-4-1-5 Arbre de décision :

Un arbre de décision est une structure simple récursive permettant d'exprimer un processus de classification séquentiel au cours duquel une correspondance est établie entre un objet décrit par un ensemble de caractéristiques (attributs), et un ensemble de classes disjointes. Chaque feuille de l'arbre dénote une classe et chaque nœud intérieur un test portant sur un ou plusieurs attributs, produisant un sous-arbre de décision pour chaque résultat possible du test [05].

I-4-2 Classification non supervisée :

I 4-2-1 Définition :

L'objectif de ces méthodes est de regrouper les individus en un nombre restreint de classes homogènes sans connaissances à priori [06].

L'apprentissage non supervisé consiste à inférer des connaissances sur des classes sur la seule base des échantillons d'apprentissage, et sans savoir *a priori* à quelles classes ils appartiennent [03].

On distingue aussi les approches de classification non hiérarchiques et les méthodes de classification hiérarchiques.

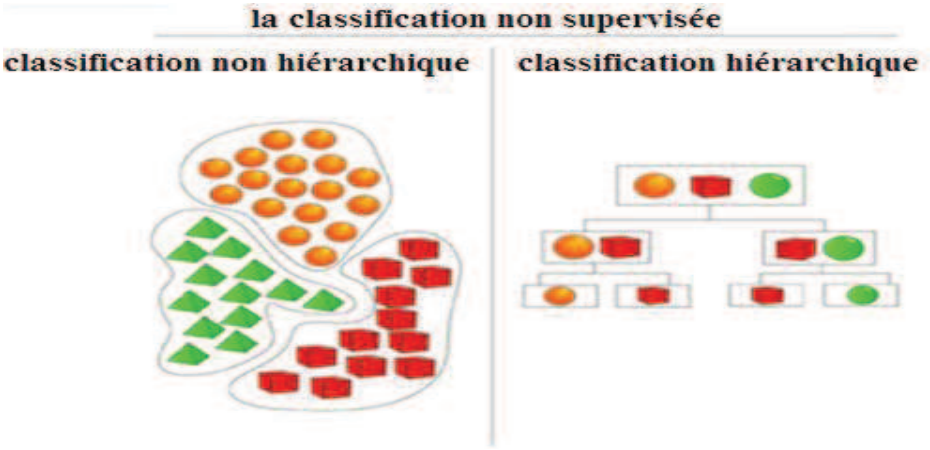

Figure I-1 : les deux types de clustering hiérarchique/non hiérarchique.

I-4-2-2 Méthodes non hiérarchiques :

Ce sont des méthodes qui produisent directement une partition en un nombre fixé de classes [05].

Regrouper n individus en k classes de telle sorte que les individus d'une même classe soient le plus semblables possible et que les classes soient bien séparées [06].

Parmi ces méthodes, nous retrouvons:

a- k-means :

Ces méthodes construisent k classes à partir d'un ensemble de n individus, tout en minimisant la quantité : $\sum_{r=1}^{k}\sum_{x_i \in c_r}(x_i - g_r)^2$

Ou :

- C_r est la classe numéro r
- X_i est un individu dans une classe

G_r est le centre de classe C_r .

L'algorithme général de ces méthodes :

Donnée : k le nombre maximum de classe désiré.

Début

(1) Choisir k individus au hasard (comme centre des classes initiales)

(2) Affecter chaque individu au centre le plus proche

(3) Recalculer le centre de chacune de ces classes

(4) Répéter l'étape (2) et (3) jusqu'à stabilité des centres

(5) Editer la partition obtenue

Fin

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

- Méthode de centre mobile
- Méthode de nuée dynamique

Ces méthodes donnent la plupart de temps une partition localement optimale, il est donc conseiller d'effectuer plusieurs exécutions et comparer les différentes résultats obtenues. Une suggestion consiste [09] à appliquer la méthode des k-means dans une première étape à plusieurs sous-ensembles de données extraits de l'ensemble total, et la meilleure partition obtenue fournit les centres à utiliser au départ de l'algorithme appliqué à l'ensemble total des données [01].

- **Nuées dynamiques (*isodata*) :**

Cette méthode a été proposée par Diday.E en 1972 **[07**]. Elle peut être considérée comme une généralisation de la méthode des centres mobiles. Le principe de la méthode est le suivant: on tire au hasard *k* noyaux parmi une famille de noyaux (chaque noyau contient un sous-ensemble d'individus). Puis chaque point de l'ensemble d'apprentissage est affecté au noyau dont il est plus proche. On obtient ainsi une partition en *k* classes dont on calcule les noyaux. On recommence le processus avec les nouveaux noyaux et ainsi de suite jusqu'à ce que la qualité de la partition ne s'améliore plus [01].

• **Algorithme**

1. Soit E l'ensemble à classer.
2. Soit f une fonction qui détermine un noyau d'une classe donnée.
3. Soit g une fonction qui détermine une classe autour d'un noyau donné *avec* **Noyau**: c'est l'ensemble d'éléments qui agit comme un centre.
4. Soit W une fonction qui mesure l'homogénéité des classes d'une partition donnée et un ensemble de noeuds donné.
5. Soit K le nombre de classes à créer.
6. Choisis K noyaux dans E.
7. Tant que W n'est pas satisfaisant
8. Utilise g pour déterminer une classe autour de chaque noyau.
9. Utilise f pour déterminer les noyaux de ces classes.
10. Fin Tant que

- **Méthodes de centres mobiles :**

Cette méthode est développée par Forgy en 1965[08]. Elle consiste à construire une partition en k classes en sélectionnant k individus comme centres de classes tirés au hasard de l'ensemble d'individus. Après cette sélection, on affecte chaque individu au centre le plus proche en créant k classes, les centres des classes seront remplacer par les centres de gravité et nouveaux classes seront crées par le même principe [01].

Donnée : k le nombre maximum de classes désiré.

Début

(1) Choisir k individus au hasard (comme centre des classes initiales)

(2) Affecter chaque individus au centre le plus proche

Ce qui donne une partition en k classes $P_1 = \{C_1, C2, \ldots\ldots, C_k\}$

(3) On calcule les centres de gravité des chacune des classes de P_1 ce qui donne K nouveaux centres de classes.

(4) Répéter l'étape (2) et (3) jusqu'à deux itérations successives donnent la même partition

(5) éditer la partition obtenue.

FIN

Généralement la partition obtenue est localement optimale car elle dépend du choix initial des centres. Pour cela les résultats entre deux exécutions de l'algorithme sont significativement variés [01].

b- K-médoïds :

Dans ces méthodes une classe est représenté par un de ces individus (médoïde),c'est une méthode itérative combinant la réaffectation des individus dans des classes avec une intervention des médoïdes et des autres individus , c'est une méthode simple parce qu'elle couvre n'importe qu'elle type de variables , quand des médoids sont choisis, des classes sont définis comme sous-ensembles des individus près des médoïdes les plus proches par rapport à une mesure de distance choisie .

Il est alors plus judicieux de choisir comme centre de groupe un individu présent dans le groupe et non un individu calculé. La médoïde d'un groupe est l'individu possédant la dissimilarité moyenne la plus faible d'avec les autres individus du groupe [01].

- ***Avantages :***

- Bonne résistance aux données erronées,
- Flexibles avec tout type de distance,

- ***Inconvénients :***

- Nécessité de spécifier le nombre de clusters k,
- Complexité de chaque itération.

En conclusion, les méthodes non hiérarchiques permettent de traiter rapidement de grands ensembles d'individus, mais elles supposent que le nombre des classes est fixé au départ. Si le nombre de classes n'est pas connu ou si ce nombre ne correspond pas à la configuration véritable de l'ensemble d'individus (d'où le risque d'obtenir des partitions de valeurs douteuses), il faut presque toujours tester diverses valeurs de *k,* ce qui augmente le temps de calcul. C'est pourquoi, lorsque le nombre des individus n'est pas trop élevé, on préfère utiliser les méthodes hiérarchiques [05].

I-4-2-3 Méthodes hiérarchiques :

La classification hiérarchique consiste à effectuer une suite de regroupements en classes de moins en moins fines en agrégeant à chaque étape les objets ou les groupes d'objets les plus proches. Elle fournit ainsi un ensemble de partitions de l'ensemble d'objets [10]. Cette approche utilise la notion de distance, qui permet de refléter l'homogénéité ou l'hétérogénéité des classes. Ainsi, on considère qu'un élément appartient à une classe s'il est plus proche de cette classe que de toutes les autres. La figure I.2 est une illustration du principe des méthodes hiérarchiques [05].

Dans cette figure, on représente la suite de partitions d'un ensemble *{a, b,* c, *d, e}* :

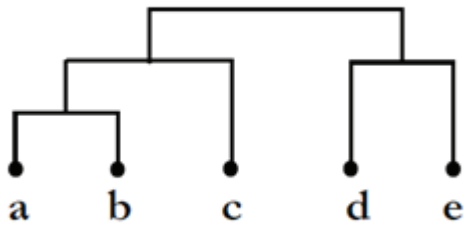

Figure I-2 : La partition hiérarchique.

a- Les critères de dissimilarité :

On peut mesurer la dissimilarité entre deux objets par [02] :

- ***La distance euclidienne*** : (aussi appelée la distance à vol d'oiseau) Un rapport de clusters analysis en psychologie de la santé a conclu que la mesure de la

distance la plus courante dans les études publiées dans ce domaine de recherche est la distance euclidienne ou la distance au carré euclidienne.

$$d^2(x_1 , x_2) = \sum_i (x_{1i} - x_{2i})^2 = (x_1 - x_2)(x_1 - x_2)'$$

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

$$d^2(x_1 , x_2) = \sum_i |x_{1i} - x_{2i}|$$

- ***La distance de Mahalanobis*** **:**

Corrige les données pour les différentes échelles et des corrélations dans les variables, L'angle entre deux vecteurs peuvent être utilisés comme mesure de distance quand le regroupement des données de haute dimension. Voir l'espace produit scalaire.

$$d^2 (x_1 , x_2) = (x_1 - x_2)C^{-1} (x_1 - x_2)$$

où (C = covariance)

- ***La distance de Hamming :***

mesure le nombre minimum de substitutions nécessaires pour changer un membre dans un autre. Elle permet ainsi , de quantifier la différence entre deux séquences de symboles, généralement utilisée dans le cas des valeurs discrètes (vecteurs)

$$d (a, b) \sum_{i=0}^{n-1} (a_i \oplus b_i)$$

Exemple : Considérons les suites binaires suivantes :

a=(0 0 0 1 1 1 1) et b= (1 1 0 1 0 1 1) alors d =1+1+0+0+1+0+0

La distance entre a et b est égale à 3 car 3 bits diffèrent.

- ***La métrique Minkowski*** : Pour les données dimensionnelles, c'est la mesure populaire

$$d_p (x_i , x_j) = (\sum_{k=1}^{d} (|x_{ik} - x_{jk}|^p)^{\frac{1}{p}}$$

où d est la dimensionnalité des données.

La distance euclidienne est un cas particulier où p = 2, alors que Manhattan p = 1. Néanmoins, il n'existe pas de directives générales théoriques pour la sélection d'une mesure à une application donnée. Une autre question, est de savoir comment mesurer la distance entre 2 classes D(C_1 ; C_2) ? Pour cela il ya certaines fonctions permettent de mesurer cette distance comme :

- plus proche voisin : min (d (i , j) , $i \in C_1$, $j \in C_2$)
- diamétre maximum : max (d (i, j) ,$i \in C_1$, $j \in C_2$)

- Distance moyenne :

$$\frac{\sum_{i,j} d(i,j)}{n_1 n_2}$$

- Distance des centres de gravité : d (μ_1, μ_2)
- Distance de Ward : $\sqrt{\frac{n_1 n_2}{n_1+n_2}}$ d (μ_1, μ_2)

b- Algorithme de la CAH:

Il en existe un bon nombre d'algorithmes pour les méthodes hiérarchiques, on se limite à présenter l'algorithme de la méthode directe CAH (Classification Ascendante Hiérarchique). Ce dernier est celui le plus simple, il est du à Lance et Wiliam (1967).

- ***Initialisation***

Construction du tableau des distances, peu importe la formule utilisée pour le construire car l'algorithme de la C.A.H est indépendant de la métrique utilisée. Ainsi, entre chaque couple de points (x, y) de M, nous disposons d'une valeur d(x, y). La partition initiale est la plus fine P_0 de M.

- ***Regroupement***

Parcourir le tableau des distance pour déterminer le couple d'élément (x^*, y^*) les plus proches :

$$d(x^*, y^*) \leq \min_{x,y \in M}\{d(x,y)\}$$

On réunit les deux éléments dans une même classe $A = x^* \cup y^*$, les autres classes restent inchangées. Nous obtenons une nouvelle partition p_i moins fine que la précédente.

- ***Tableau des distances***

La classe A sera vue comme un seul point. Il faut donc calculer les distances qu'il y a entre le point A, qui est un ensemble de cardinal supérieur à un, et tous les autres points qui ne sont pas dans A et qui peuvent être des singleton. Par souci de généralité, nous les notons B.

$$d(A,B) \; ; \; B \not\subseteq A$$

Pour cela, on peut utiliser l'un des six critères de dissimilarité proposés plus haut.

Nous disposons alors d'un nouveau tableau des distances ayant une ligne et une colonne de moins que le précédent dont il ne diffère que par la ligne et la colonne qui correspond au point A.

- ***Condition d'arrêt***

Si nous avons atteint la partition du niveau souhaité, généralement c'est la partition grossière, celle qui ne comporte qu'une seule classe réunissant la totalité des points, alors, c'est terminé. Dans le cas contraire, nous repartons de l'étape « Regroupement » à partir du tableau des distances calculé à la suite du précédent regroupement.

FIN

Malgré le nombre important des méthodes de classification, il n'est existe pas un algorithme qui répond à tout les demandes et plusieurs problématiques restent encore ouverte dans le cadre de la classification.

I-5 Conclusion :

Plusieurs méthodes sont proposées pour le problème général de la classification. Ils différent par les mesures de proximités qu'ils utilisent ,la nature des données qu'ils traitent et l'objectifs finale de la classification ,chacune de ces méthodes possède ses points forts et ses points faibles, les méthodes hiérarchiques ascendantes sont utilisées en cas des données de petite taille car la complexité est très élevée , et Si des problèmes de temps d'exécution se posent, alors c'est les méthodes des K-means qui sont utilisées. Pour cela on s'est intéressé à cette dernière qui sera détaillé dans le chapitre suivent.

Chapitre II : k-means

II-1 introduction :

En quelques mots, la classification automatique est la tâche qui consiste à regrouper, de façon non supervisée , un ensemble d'objets ou plus largement de données, de telle manière que les objets d'un même groupe (appelé cluster) sont plus proches (au sens d'un critère de (dis)similarité choisi) les unes au autres que celles des autres groupes (clusters). Il s'agit d'une tâche principale dans la fouille exploratoire de données, et une technique d'analyse statistique des données très utilisée dans de nombreux domaines, y compris l'apprentissage automatique, la reconnaissance de formes, le traitement de signal et d'images, la recherche d'information, etc.

L'idée est donc de découvrir des groupes au sein des données, de façon automatique [11].

Dans ce cadre plusieurs méthodes ont été développées, la plus populaire est celle des k moyennes (K-means), elle doit sa popularité à sa simplicité et sa capacité de traiter de larges ensembles de données [12].

II -2 Définition :

L'algorithme k-means mis au point par McQueen en 1967[13], un des plus simples algorithmes d'apprentissage non supervisé , appelée algorithme des centres mobiles [14] [15] , il attribue chaque point dans un cluster dont le centre (centroïde) est le plus proche. Le centre est la moyenne de tous les points dans le cluster , ses coordonnées sont la moyenne arithmétique pour chaque dimension séparément de tous les Points dans le cluster c'est à dire chaque cluster est représentée par son centre de gravité.

II-3 Exemples d'applications :

- **Marketing** : segmentation du marché en découvrant des groupes de clients distincts à partir de bases de données d'achats.
- **Environnement** : identification des zones terrestres similaires (en termes d'utilisation) dans une base de données d'observation de la terre.
- **Assurance** : identification de groupes d'assurés distincts associés à un nombre important de déclarations.
- **Planification de villes** : identification de groupes d'habitations suivant le type d'habitation, valeur, localisation géographique, … [03].

II-4 Algorithme kmeans :

Entrée

Ensemble de N données, noté par x

Nombre de groupes souhaité, noté par k

Sortie

Une partition de K groupes $\{C_1, C_2, ...Ck\}$

Début

1) Initialisation aléatoire des centres C*k* ;

Répéter

2) Affectation : générer une nouvelle partition en assignant chaque objet au groupe dont le centre est le plus proche :

$$x_i \in c_k \text{ si } \forall_j |x_i - \mu_k| = \min |x_i - \mu_j| \qquad (1)$$

Avec μ_k le centre de la classe K ;

3) Représentation : Calculer les centres associe à la nouvelle partition ;

$$\mu_k = \frac{1}{N} \sum_{x \in c_k} x_i \qquad (2)$$

Jusqu'à convergence de l'algorithme vers une partition stable *;*

Fin.

La principale limite de cette méthode est la dépendance des résultats des valeurs de départ (centres initiaux). À chaque initialisation correspond une solution différente (optimum local) qui peut dans certain cas être très loin de la solution optimale (optimum global). Une solution naïve à ce problème consiste à lancer l'algorithme plusieurs fois avec différentes initialisations et retenir le meilleur regroupement trouvé. L'usage de cette solution reste limité du fait de son coût et que l'on peut trouver une meilleure partition en une seule exécution [10] .

II- 5 Les différentes versions de k-means :

II-5-1 Global k-means :

Global k-means [15] est une solution au problème d'initialisation du k-means, elle est fondé sur les données et vise à atteindre une solution globalement optimale. Elle consiste à effectuer un clustering incrémental et à ajouter dynamiquement un nouveau centre suivi par l'application du k-means jusqu'à la convergence.

Les centres sont choisis un par un de la façon suivante : le premier centre est le centre de gravité de l'ensemble des données (résultat de l'application du k-means avec k=1), les autres centres sont tirés de l'ensemble de données ou chaque donnée est une candidate pour devenir un centre, cette dernière sera testée avec le reste de l'ensemble, le meilleur candidat est celui qui minimise la fonction objectif

$$\mathbf{J} = \sum_{i=1}^{k} \sum_{x_j \in c_i} \left\| x_j - c_i \right\|^2$$

L'algorithme suivant permet d'illustrer le principe :

Entrée

Ensemble de N données, notés par x ;

Nombre de groupes souhaiter, noté par k ;

Sortie

Une partition de K groupes { C_1 ,C_2 ,...C_k }

Début

1) C_1 = Centre de gravité de l'ensemble des données ;

Répéter

2) Initialiser les centres i-1 par le résultat de l'étape précédente ;

3) Trouver l'iéme centre :

Pour chaque donnée x **faire**

3.1) Considère x comme étant le iéme centre ;

3.2) Affecter les données aux plus proche centre ;

3.3) Calculer l'erreur quadratique pour Ci=x ;

$$\mathbf{J} = \sum_{i=1}^{k} \sum_{x_j \in c_i} \left\| x_j - c_i \right\|^2$$

Fin faire

3.4) Garder le centre Ci = x qui minimise

l'erreur quadratique ;

4) Appliquer le k-means jusqu'à la convergence ;

Jusqu'à obtenir une partition en k groupes ;

II-5-2 Initialisation par le mal classé :

II-5-2-1 Principe :

L'absence d'un signe indiquant si l'optimum global est atteint ou pas fait penser à la possibilité d'améliorer les résultats.

Observant l'équation : $\mathbf{x_i} \in \boldsymbol{c_k}$ **si** $\forall_j |\mathbf{x_i} - \boldsymbol{\mu_k}| = \mathbf{min} \left|\mathbf{x_i} - \boldsymbol{\mu_j}\right|$ (voir dans l'algorithme de k-means un peu plus haut**),** un objet est affecté à un groupe s'il lui est le plus proche, plus la distance diminue plus la probabilité d'appartenance à ce groupe augmente, dans le cas contraire, l'objet le plus loin de son groupe d'appartenance est considéré comme étant mal classé, il fera certainement un bon candidat afin de former le nouveau centre.

Le global k-means est amorcé par un seul groupe ayant pour représentant le centre de gravité de l'ensemble des données, dans certain cas, cette partie de l'espace est vide (figure 1) ce qui permet de dégrader la classification, nous proposons d'amorcer l'initialisation du k-means avec deux groupes, les centres de ces groupes doivent assurer la séparabilité des données au cours de classification, il est évidant de choisir les deux données les plus éloignées [16].

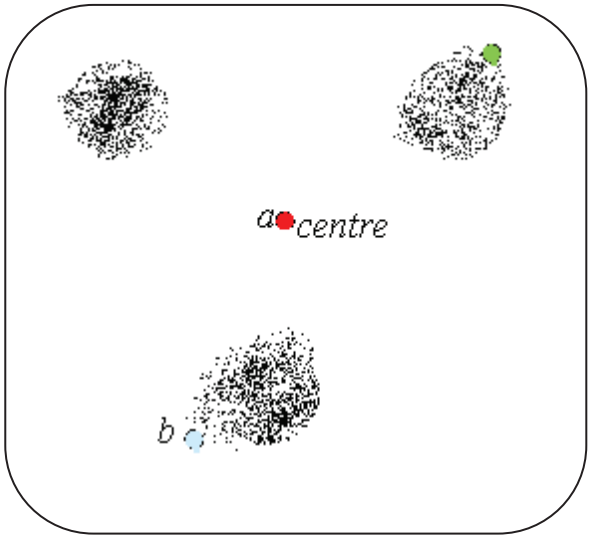

Figure II-1 classification par le principe d' initialisation par le mal classé

(a) Le centre des données en rouge,
(b) le bleu et le vert représentent les deux objets les plus éloigné

II-5-2-2 Algorithme :

Début

1) Création d'une matrice de distance

2) Choisir les deux éléments les plus éloignés (ils représentent les deux premiers centres) ;

TANT QUE le nombre de classes souhaité n'est pas atteint **Faire**

3) Affecter les individus aux noyaux disponibles ;

4) Sélectionner un élément mal classé (celui qui possède la plus grande distance de son centre le plus proche) ;

5) Ajouter cet individu à l'ensemble des noyaux ;

6) Augmenter le nombre des noyaux ;

Fin TANTQUE

Fin

II-5-3 L'approche incrémental (ou Modified Fast Global K-means) :

Cette approche incrémental de classification est similaire à celle du globale k-means, la différence entre elles réside dans les points suivant :

- Le nombre de points initiaux, dans notre cas deux au lieu de un seul dans le global k-means.
- La recherche du nouveau centre ce limite à la recherche de l'élément le mal classé au lieu de testé toutes les données.

II-5-3-1 Algorithme :

Entrée

Ensemble de N données, notés par x ;

Nombre de groupes souhaiter, noté par k ;

Sortie

Une partition de K groupes $\{C_1, C_2, ...Ck\}$

Début

1) $c_1 = x_1$;

$c_2 = x_2$; avec $d(x_1, x_2) = \max_{i,j \in \{1,..,N\}, i \neq j} (d(x_i - x_j))$

Répéter

2) Initialiser les centres i-1 par le résultat de l'étape précédente ;

3) Trouver l'i^{éme} centre Ci:

$$C_i = x : x = \max_{i \in [1,n]} (d_{k-1}^i)$$

Avec d_{k-1}^i la distance entre x_i et son plus proche centre parmi les k-1 centre

4) Appliquer le k-means jusqu'à la convergence ;

Jusqu'à obtenir une partition en k groupes ;

Fin.

Grâce au faible cout de la stratégie de choix du nouveau centre, il est clair que l'approche proposée est plus rapide que le global k-means [17].

II-6 Avantages de k-means :

Nous pouvons citer quelque avantages de k-means par :

- L'avantage de ces algorithmes est avant tout leur grande simplicité.
- Tend à réduire l'erreur quadratique.
- Applicable à des données de grandes tailles [09].

II-7 Inconvénients :

- Le nombre de classe doit être fixé au départ.
- Ne détecte pas les données bruitées.
- Le résultat dépend de tirage initial des centres des classes.
- Les clusters sont construits par rapports à des objets inexistants (les milieux)
- N'est pas applicable en présence d'attributs qui ne sont pas du type intervalle [09].

II-8 Conclusion :

L'algorithme du K-Means fut l'une des approches non-hiérarchiques les plus populaires. Il est l'outil de classification le plus utilisé dans les applications scientifiques et industrielles. Quoiqu'il ne fonctionne pas très bien pour les attributs catégoriels, il a un bon sens géométrique et statistique pour les attributs numériques.
Pour cette raison on a choisit la version incrémental pour appliquer le K means dans notre application qui sera détaillé dans le chapitre suivant.

Chapitre III : Application

III-1 Préliminaire :

Dans ce dernier chapitre et après l'aperçu théorique des chapitres précédents, nous présentons le côté pratique de notre application .Notre but est de classifier des points d'une manière automatique sans intervention ou connaissances préalables en utilisant comme algorithme de classification « approche incrémentale » parce que cette technique produit des groupes homogènes en un temps très réduit par rapport au autres version de k-means.

Nous commençons par la description de la base utilisée, le choix de l'environnement de travail ainsi que les étapes fondamentaux de la conception de notre application. Celle-ci porte le nom ≪ KMEANS ≫.

III-2 Le système d'exploitation :

L'environnement WINDOWS XP a été choisi comme environnement de travail pour notre logiciel pour les raisons suivantes :

- Une très bonne gestion de mémoire ;
- Une architecture orientée évènement ;
- Un graphisme indépendant des périphériques ;
- La notion de ressources.

III-3 Langage de programmation :

C++Builder est un **IDE**, un environnement de développement intégré. Il regroupe tout un ensemble d'outils permettant d'effectuer un maximum de tâches de développement au sein du même environnement de travail.

C++ Builder est de plus un environnement de développement visuel C++ RAD (Rapid Application Development). Il permet de construire rapidement des applications en utilisant des composants et simplifie au maximum l'écriture du code et la réalisation de l'interface. On peut ainsi très rapidement se consacrer à la partie "métier" du code (le code réellement utile de l'application).

C++Builder permet également le développement rapide d'applications base de données, ainsi que des applications-serveurs web [17].

Figure III.1 : Le c++ builder .

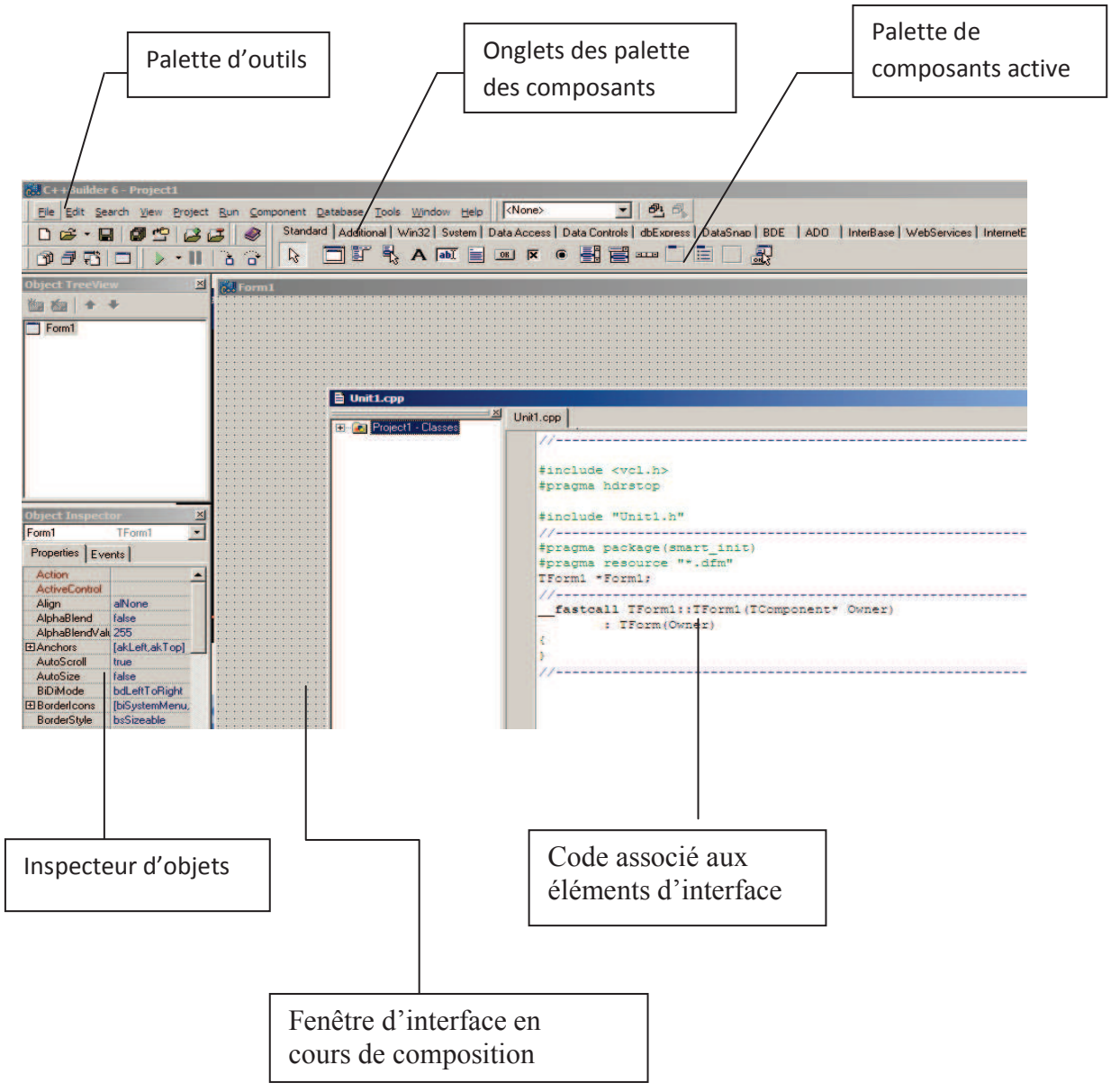

Figure III.2 : L'interface de C++ Builder

III-4 Conception :

III-4-1 Organigramme de l'algorithme de k-means :

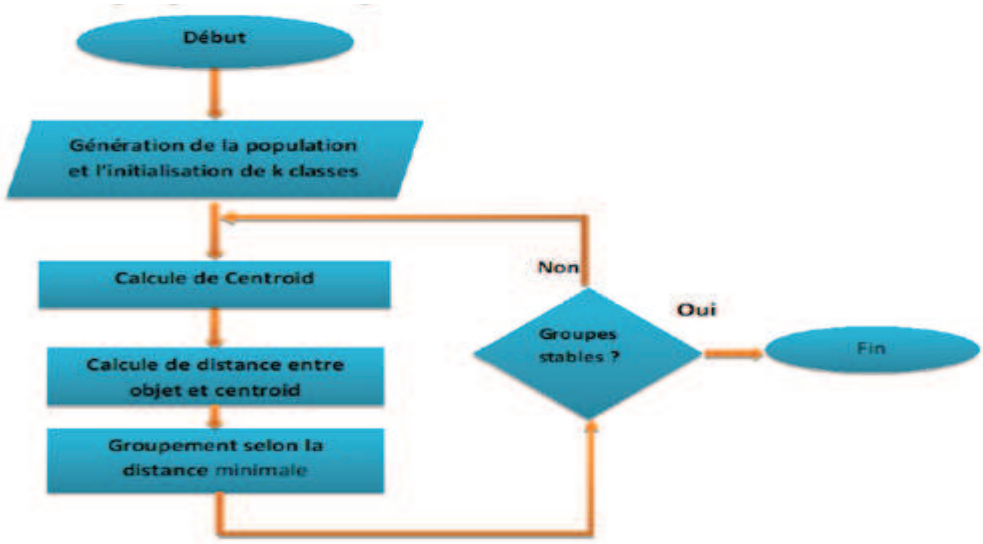

Figure III.3 : organigramme de l'algorithme k-means

III-4-2 Diagramme de cas d'utilisation de k-means :

L'utilisateur étant l'acteur principal. Les cas d'utilisation de base qui vont être mis en évidence pour réaliser l'ensemble des groupes seront :

- Configuration de k-means.
- Saisir le nombre de points.
- Saisir le nombre cluster.
- Clutering.

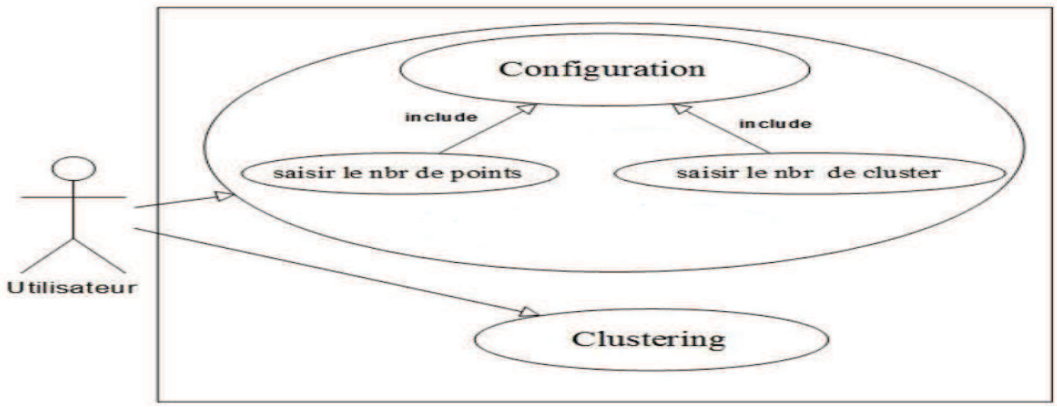

Figure III.4 : Diagramme de cas d'utilisation pour l'algorithme de k-means

III-5 Description de l'application :

III-5-1 Interface et composants :

En cliquant sur « classification» un menu déroulant apparaît vous permettant d'effectuer les fonctions suivantes :

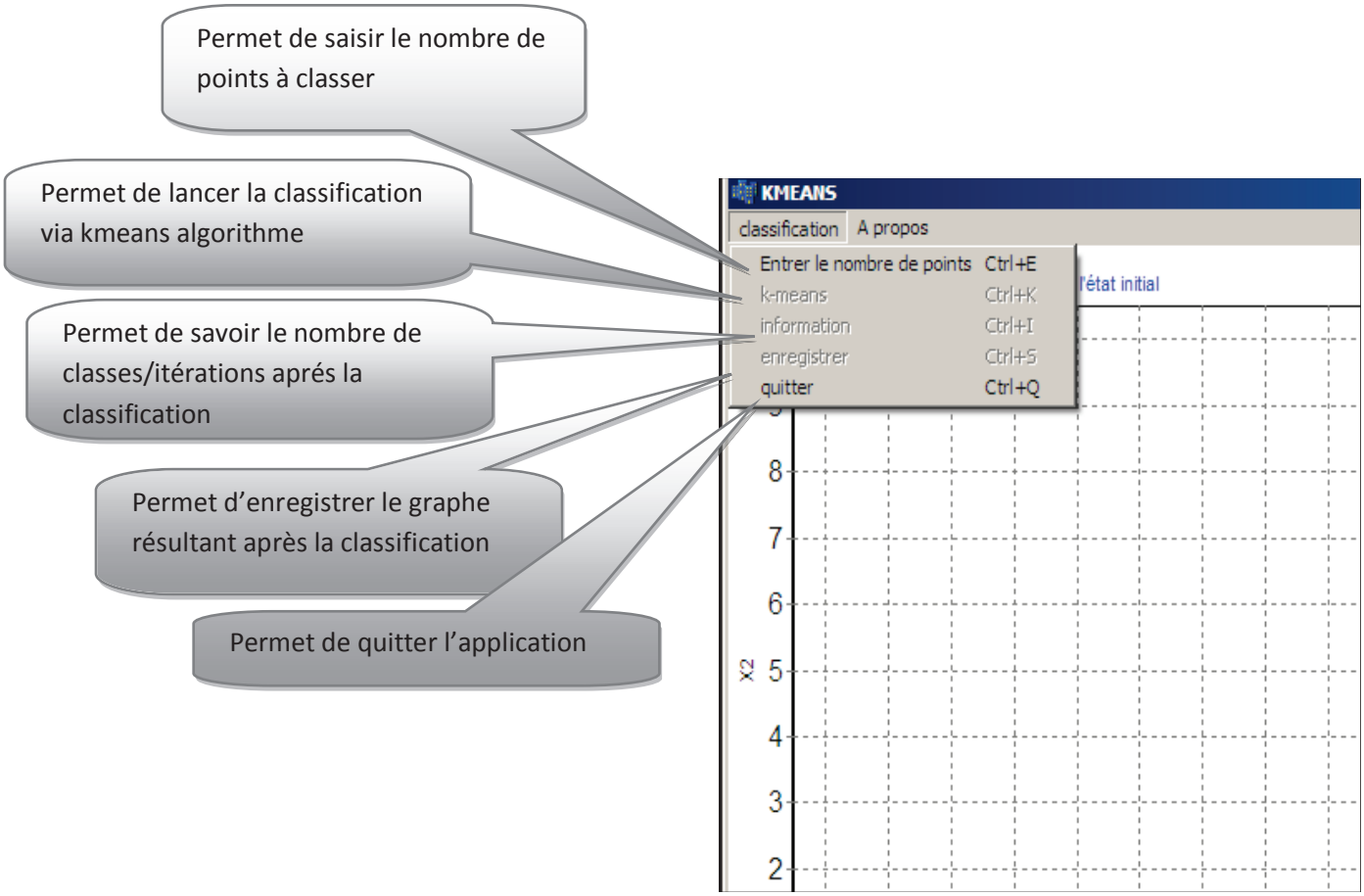

Figure III.5 : Menu classification

III-5-2 Les étapes de démarche :

a- L'étape (1) :

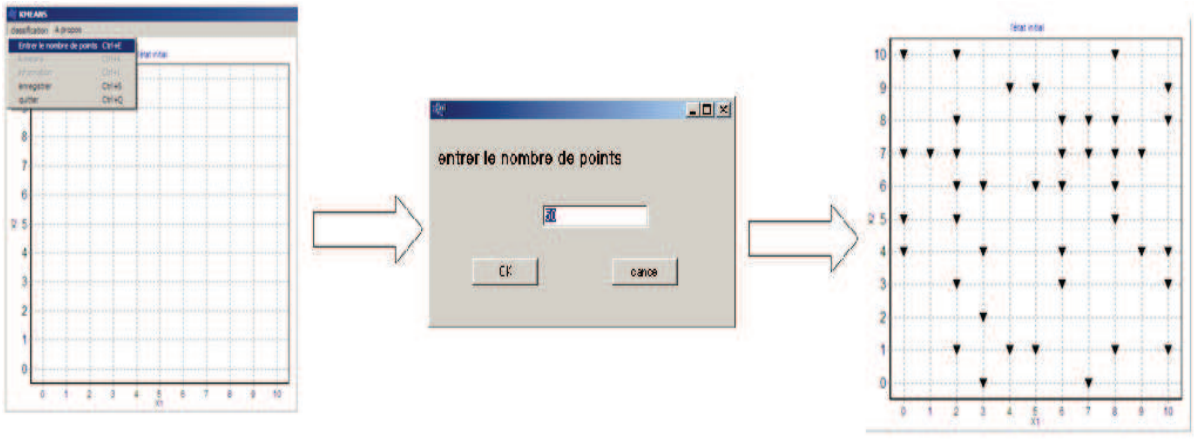

Figure III.6 : **Saisir le nombre de points**

b- L'étape (2) :

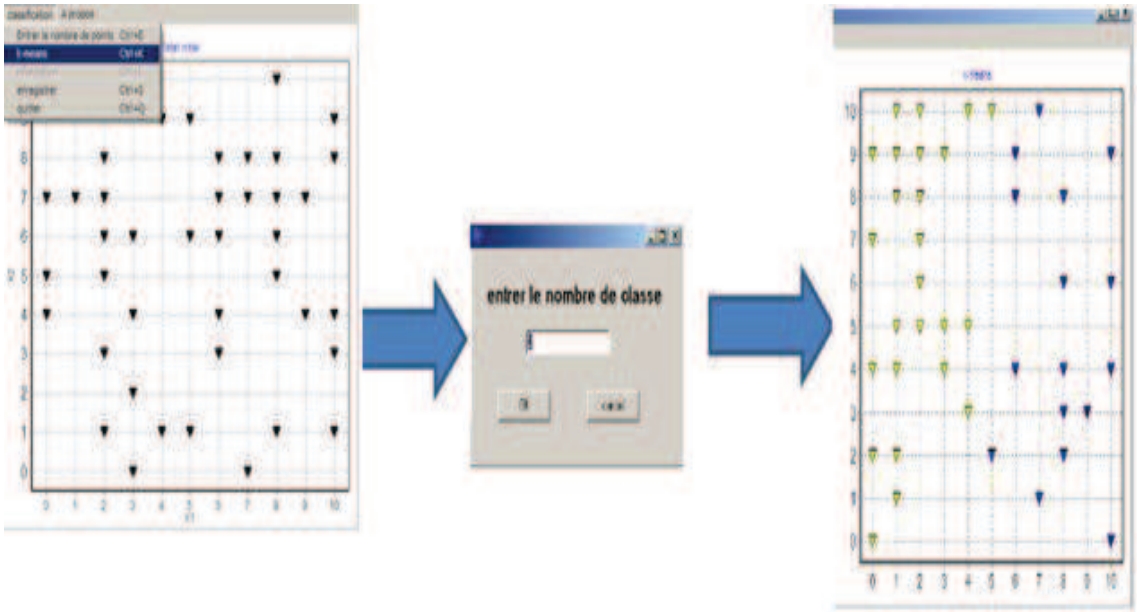

Figure III.7: Classification k-means

c- L'étape (3) :

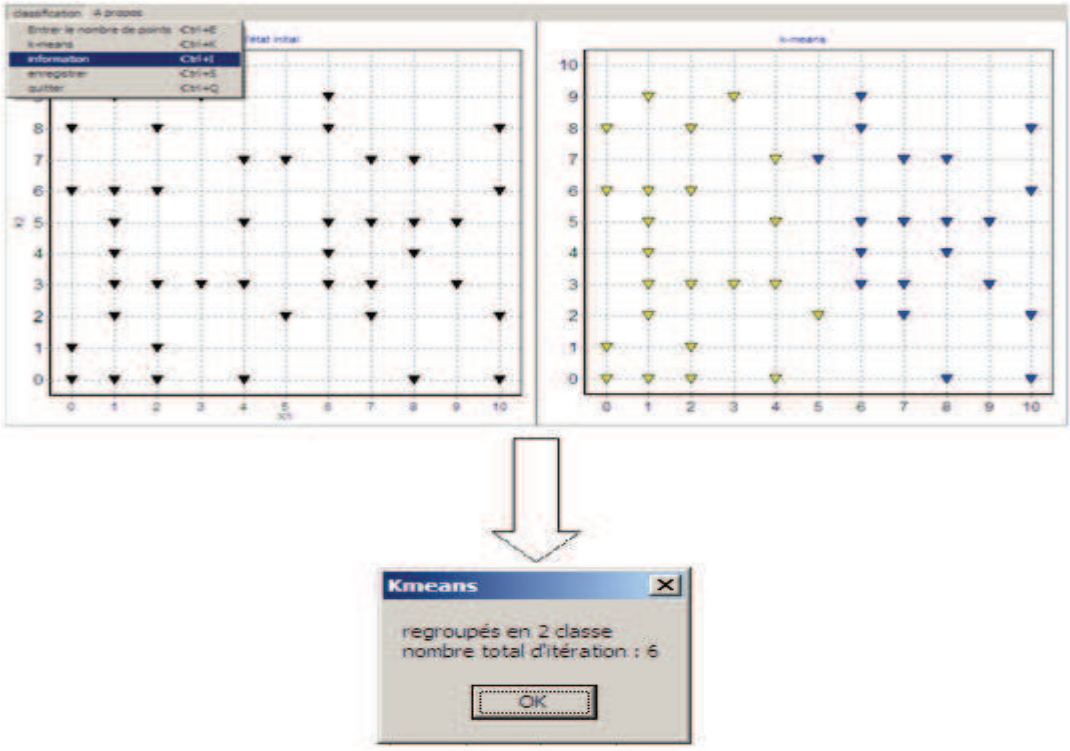

Figure III.8 : Des informations sur la classification

d- l'étape(4) :

Enregistrer le graphe résultant de classification dans un chemin choisis comme image de format (.bmp) (figure III.11)

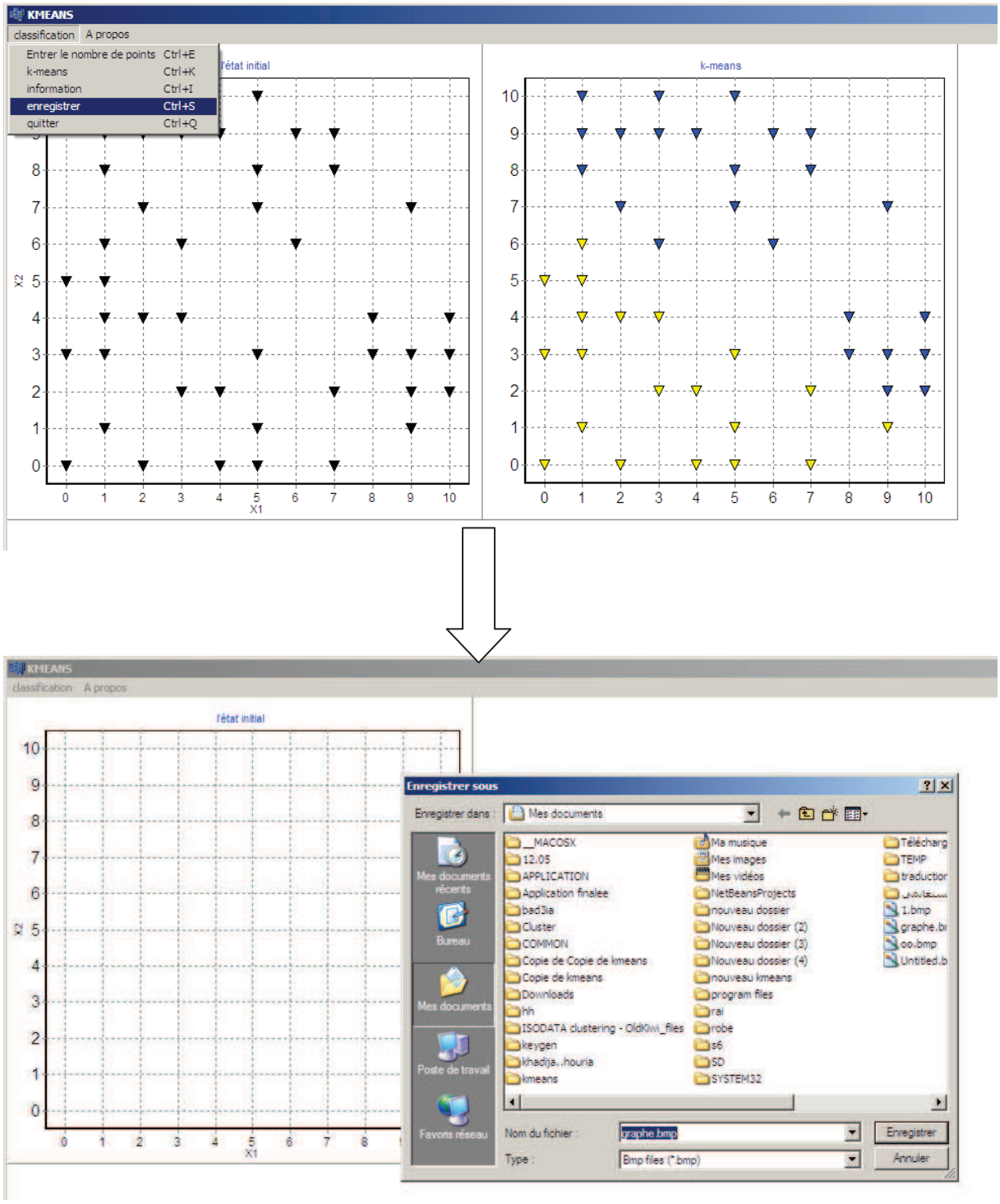

Figure III.9 : Enregistrer

e- L'étape (5):

On peut aussi initialiser tout les points pour faire une autre classification avec des nouvelles données.

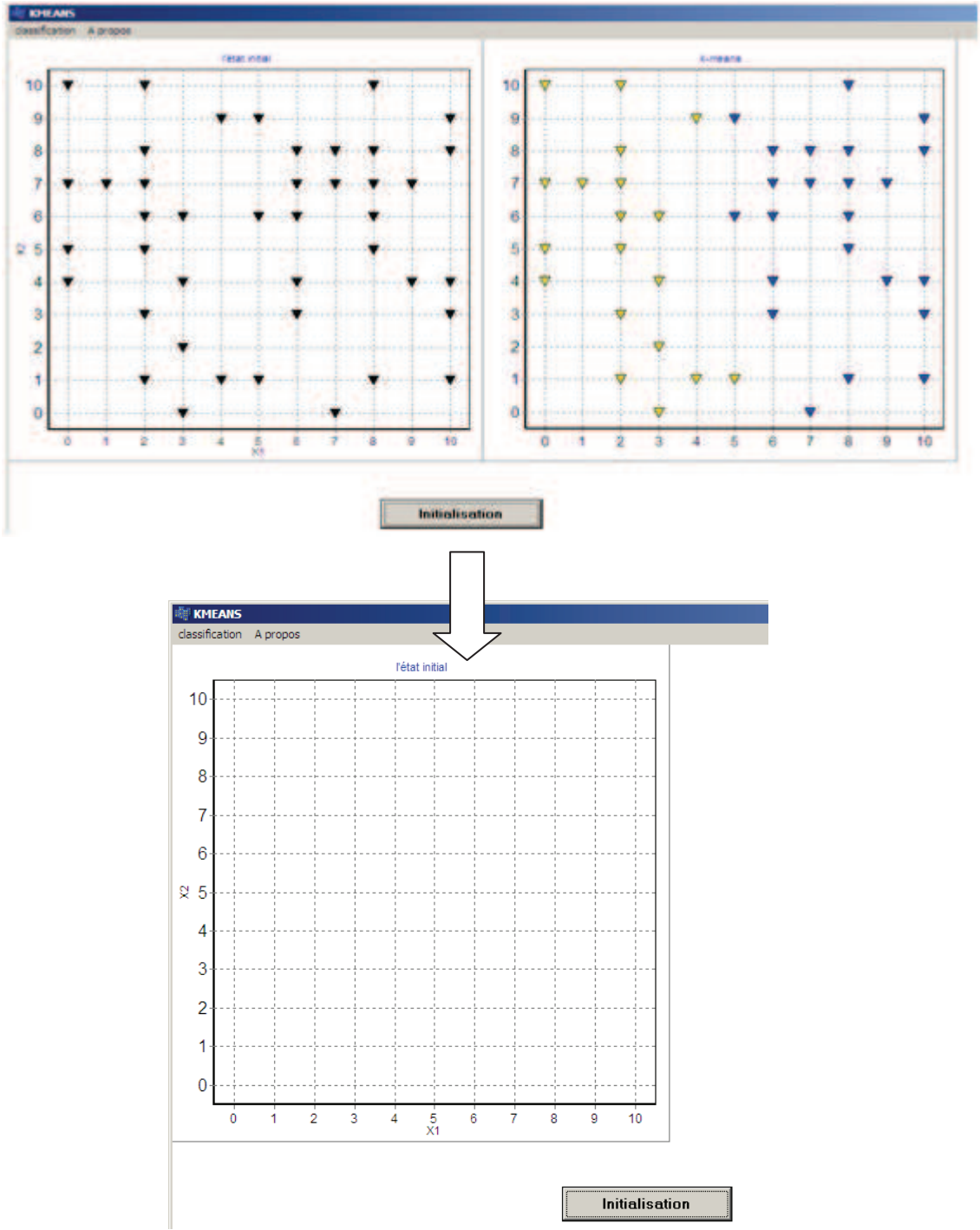

Figure III.10 : **Initialisation**

III-4-3 Exemples :

Notre programme a pour but de réaliser une application de classification k-means. Ce programme se caractérise par son graphique simple qui facilite son utilisation.

Notre programme permet de faire le la classification à partir d'un ensemble de points avec des coordonnée choisis aléatoirement dans un plan à deux dimension.

a. Exemple1 :

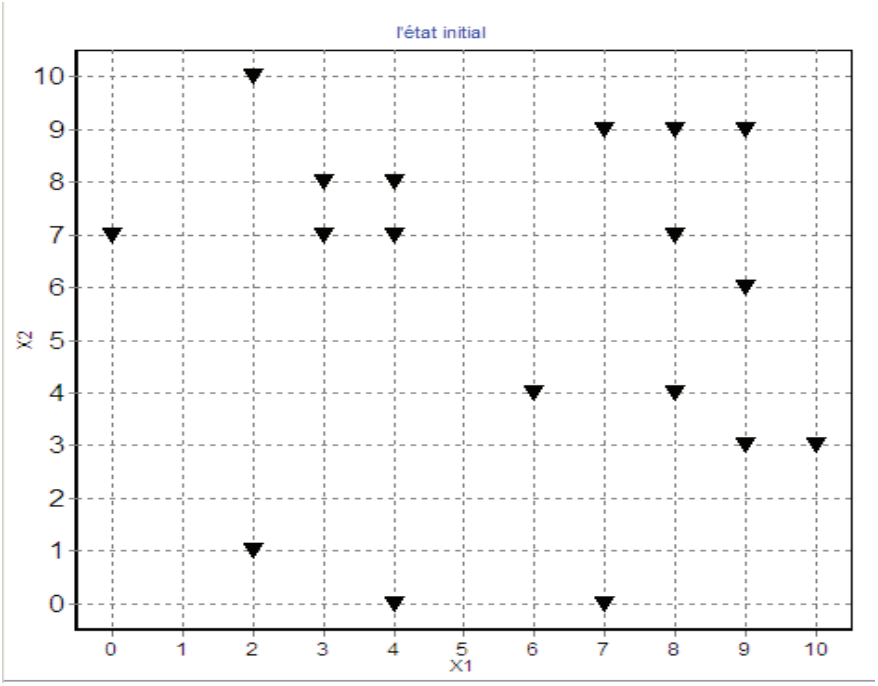

Figure III.11 :l'état initial avec 18 points

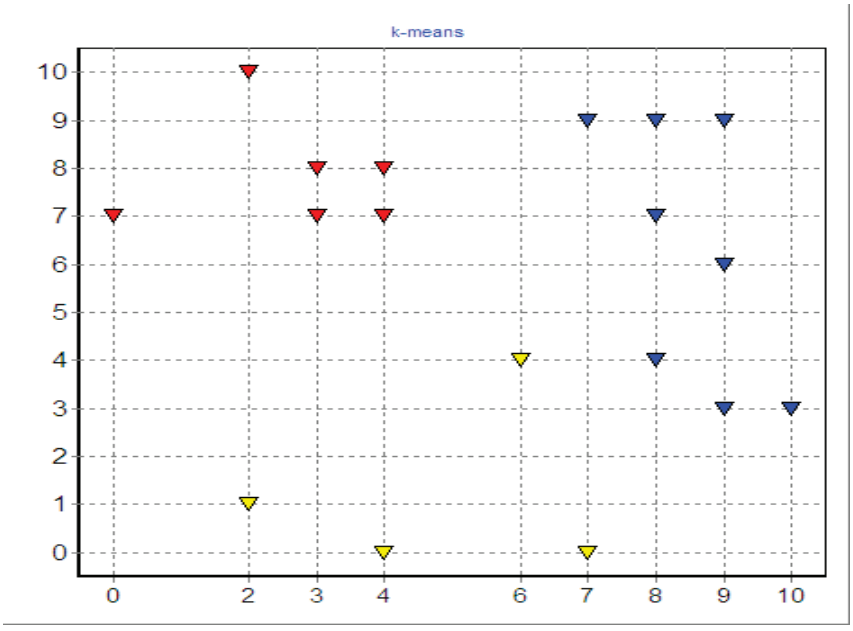

Figure III.12 : le résultat de classification en 3 classes en 2 itérations

b. Exemple2 :

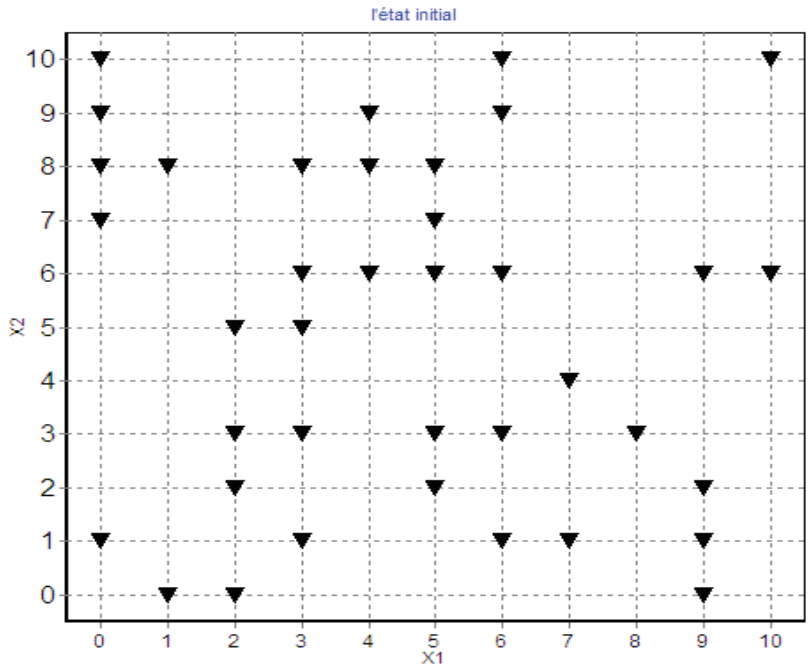

Figure III.13 : l'état initial avec 40 points

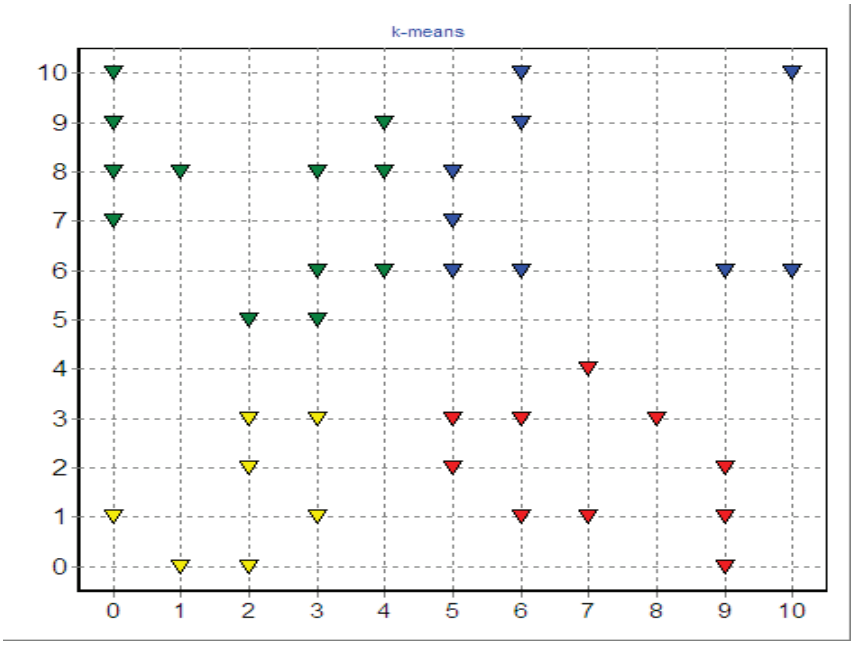

Figure III.14 : le résultat de classification en 4 classes en 3 itérations

III-6 Conclusion :

Dans ce chapitre on a appliqué l'algorithme k-means qui permet de classer un ensemble de points en des clusters homogènes , en se basant sur l'algorithme d'approche incrémentale .

Conclusion générale :

Dans plusieurs domaines des sciences sociales, nous sommes amenés à constituer des groupes homogènes en leur sein et qui diffèrent suffisamment l'un de l'autre. C'est l'objet des méthodes de classification dont fait partie la méthode des k-means, cet algorithme est une version améliorée et randomisée de la méthode des nuées dynamiques. Il est actuellement l'un des plus utilisés et des plus efficaces en analyse des données. De fait, il permet de partitionner une population finie d'éléments en un nombre K (entier) de classes homogènes.

Il est utile de noter que l'algorithme k-means est très performant en termes de temps d'exécution, mais il souffre du problème de dépendance des résultats aux choix effectués lors de l'initialisation.

On peut élargir notre travail, en essayant de comparer nos résultats avec d'autres versions de K-means, travailler sur d'autres algorithmes de classification non supervisé, et même le supervisé.

Bibliographie :

[01] Mounzer BOUBOU : "*contribution aux méthodes de classification non supervisée via des approches prétopologiques et d'agrégations d'opinion''*, thèse de doctorat, université Claud Bernard –Lyon1, 2007.

[02] KOUDRI MOHAMMED *'' modèle de mélange gaussien : application sur les images cytologique*'' Master , université Abou bakr belkaid Tlemcen ,2011

[03] Belhabib Abdelkader ,Lagha Omar : '' *Développement d'une application à base de l'algorithme de classification k-means''*, Mémoire de fin d'études pour l'obtention du diplôme de Licence en Informatique, université Abou bakr belkaid, Tlemcen ,2012

[04] Berrani, S.-A., Amsaleg, L., & Gros, P. « Recherche par similarités dans les bases de données multidimensionnelles : panorama des techniques d'indexation. » Ingénierie des systèmes d'information (RSTI série ISI-NIS), 7(5-6), pp 65-90.2002.

[05] Lamri Laoumer : ''Approche exploratoire sur la classification appliquée aux images'', Mémoire , Université du Québec à Trois-Rivières , Avril 2006

[06] Licence Professionnelle Géomatique et Environnement : ''*TRAITEMENT NUMÉRIQUE DES IMAGES Classifications non supervisées*''

[07] Diday .E, " *Optimisation en classification automatique et reconnaissance de formes".* Note Scient. IRIA nO 6, 1972.

[08] E,W Forgy cluster analysis of multivariate data : efficency versus interpretability of classification (abstract) ,Biometrics 21:768-769 ,1968

[09] S . P . Bradley ,U.M. Fayyad ,and C. Reina . Scaling clustering algorithms to large databases. In knowledge Discovery and Data Mining ,pages 9-15,1998.

[10] Cel eux .G, Diday .E, Govaert .G, " *Classification automatique de données environnement statistique et informatique".* Dunod, Informatique, 1989.

[11] Faicel CHAMROUKHI, *Classification supervisée : Analyse discriminante*, Licence 2 Sciences Pour l'Ingénieur,Université du Sud Toulon – Var, 2013.

[12] Jacob Kogan, Introduction *to Clustering Large and High-Dimensional Data*, Cambridge University Press, Cambridge, 2007.

[13] J. B. MacQueen (1967) : "*Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*", Berkeley, University of California Press,no 1, pp281-297.1967.

[14] Benzécri J.P. *L'analyse des données*. Dunod, Paris, 197,1973

[15] Likas A., Vlassis M. & Verbeek J., ‘‘*he global k-means clustering algorithm, Pattern Recognition’’*, 36, pp. 451-461.,2003

[16] Z.Guellil et L.Zaoui , ‘‘*Proposition d'une solution au problème d'initialisation cas du K-means’’*, Université des sciences et de la technologie d'Oran MB .

Web graphie :

[17] http://cpp.developpez.com/faq/bcb/?page=Le-logiciel-Cplusplus-Builder