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

Mémoire de fin d'études

pour l'obtention du diplôme de Master en Informatique

Option: *Système d'Information et de Connaissances (S.I.C)*

Thème

LE CLUSTERING PAR APPROCHE THEORIE DES JEUX

Réalisé par :

- **Dib Fatima Zohra**
- **Madani aicha**

Présenté le 03 Juillet 2018 devant le jury composé de

- *Mr. CHOUITI Sid Mohammed* *Président*
- *Mme CHAOUCHE RAMDANE Lamia.* *Encadreur*
- *Mme EL YEBDRI Zeyneb* *Examinatrice*

Année universitaire : 2017-2018

Remerciements

Avant tout nous remercions DIEU tout puissant pour l'accomplissement de ce travail, car sans DIEU rien n'aura lieu.

Nous tenons à remarcier tout d'abord notre Encadreur Madame CHAOUCHE RAMDANE Lamia, de nous avoir montré l'intérêt, mais aussi l'exigence de la recherche en informatique. Merci pour sa disponibilité, sa patience et la pertinence de ses remarques tout au long de la réalisation de ce mémoire.

Nos remerciements vont également à Mr. CHOUITI et Mme EL YEBDRI d'avoir accepté de faire partie du jury.

Nous remercions tous nos professeurs, amis et collègues pour l'ambiance amicale et fraternelle qui a régné entre nous durant toutes nos études et tous ceux qui ont contribué de près ou de loin à l'aboutissement de ce travail.

Nous tenons à remercier ainsi toute l'équipe pédagogique de l'Université de Tlemcen pour la qualité des cours ainsi que les conditions de travail et d'encadrement au sein du département d'informatique m'auront été très bénéfiques.

Nous exprimons notre profonde gratitude à nos familles pour leurs encouragements et tous leurs apports qui n'étaient pas moindre.

Dédicaces

je dédie ce modeste travail en particulier a mes adorables parents

pour tous leurs efforts et leurs sacrifices:

A ma mère amina qui m'a donné l'amour et l'affection

A mon père aissa qui m'a donné le courage et la volonté.

A mon frères : mnawer

A mes sœurs :khadija & sihem

A mes meilleurs amies dib fatéma & hadjer

A mes très chères amies Imene meddahi, hayet & imene kaid slimen & nora

A mon mari et ma belle fille rahef iness

A mes grands parents, mes oncles ,mes tantes et toute ma famille chacun avec son nom,

La liste est très longue que même la taille d'un mémoire ne suffira pas donc a tous ceux ou celle avec qui j'ai partager un moment de ma vie.

Que je dédie ce travail

Madani aicha

Dédicace :

A mes chères parents mohamed et fatima et mon chère mari abdelkarim que je dédie ce travail pour leurs bien vaillance et leurs soutiens durant tout mon cursus.

A mes chères enfants d'aujourd'hui avenir de demain abdelmoudjib, zineb, ayoub et mahmmed

Sons oublier mes chères sœurs lamia et son mari abderezzak, fatiha et son mari abdelazziz ainsi que rachida.

A tous nos enfants manel, malek, hannane, allaa, oussama, mouad, annes, sohaib et aimededdinee. à toute la famille DIB et BELABDELLI

A mes chères amis et mes collègues de travail : yahia, nacima , mohamed, saliha et ma voisine nezha

« Que je dédie ce travail

Fatima zohra.D

Sommaire

Sommaire

*

liste des figures :

liste des tableaux:

Introduction Générale

Introduction générale

Dans la vie, on fait appel souvent à la notion de groupes quand il s'agit de structurer, d'organiser ou de résumer un ensemble d'éléments comme par exemple des familles de plantes, des genres de musique, des groupes de produits, des groupes sanguins, etc… En fait un groupe est définit comme un ensemble d'éléments qui sont rassemblés en raison d'une relation particulière entre eux.

La problématique consistant à former de tel groupes de manière automatique se pose dans de nombreux domaines.

Deux approches existent pour former de tels groupes de manière automatique, la première approche intervient quand les groupes existent à priori. Elle crée un modèle qui assigne des éléments à ces groupes dans le but de découvrir un model de groupes qui généralise un ensemble de données plus large. Cette tache est appelé classification supervisé et les groupes sont appelé les classes.

La seconde émerge des groupes au sein d'un ensemble d'élément sans aucune information à priori afin de découvrir la structure sous jacente des données pour extraire de l'information. Cette tache est appelée selon les domaines classification non supervisé ou encore *clustering*. Les groupes crées sont appelés *clusters*.

En effet, Les objets dont les classes est inconnue sont généralement disponible en grand nombre et les approches non supervisées sont massivement utiliser pour traiter des données automatiquement. Cependant, de nombreux problèmes se posent lors du choix de l'algorithme de classification non supervisé à appliquer, dont on peut citer les deux plus importants :

1. L'existence d'un grand nombre de méthodes différents de clustering, mais aussi le choix de leurs paramètres qui est liés au domaine d'application
2. L'augmentation de la complexité des données qui pousse les approches non supervisé à monter leurs limites.

Notre objectif dans ce mémoire, est de proposer deux techniques de clustering par partitionnement à base de vote et à base de la théorie des jeux. La théorie des jeux peut être utilisée pour modéliser un système sous forme de jeu, règles, et joueurs. Dans notre approche, nous avons introduit l'indice de validité comme critère d'évaluation d'homogénéité. Nous cherchons à trouver la meilleure distribution de données dans les classes, où chaque donnée appartient au cluster le plus similaire. Cette distribution est appelée équilibre de Nash qui consiste à trouver une situation où chaque objet n'a pas d'intérêt à changer de cluster.

Pour cela, nous avons organisé notre mémoire en trois chapitres. Après une introduction générale, nous commençons naturellement par présenter dans le premier chapitre les différentes approches de classification supervisée et nous focalisons notre attention sur les méthodes de clustering. Puis, nous abordons la théorie des jeux évolutionnaire et son

application dans le domaine du clustering. Notre contribution ainsi que les expérimentations menées dans ce cadre sont présentées dans le dernier chapitre. Enfin, on termine par une conclusion générale et des perspectives.

I.1)Introduction

« Le seul moyen de faire une méthode instructive et naturelle, est de mettre ensemble les choses qui se ressemblent et de séparer celles qui différent les unes des autres. »

M. Georges Buffon, Histoire naturelle, 1749.

La classification est une opération de structuration qui vise à organiser un ensemble d'observation en groupes homogènes et contrastés afin de faciliter l'analyse des informations et d'effectuer des prédictions. Elle consiste aussi à regrouper les données similaires dans les mêmes classes en assurant la dé-similarité des données de classes différentes.

Deux types approches existent :

- ✓ la première approche consiste à faire émerger des groupes au sein d'un ensemble d'éléments sans aucune information à priori. Dans ce cas, cette tâche est appelée, classification non supervisée, classification automatique ou encore *clustering*.

- ✓ La seconde approche intervient quand les groupes existent à priori, et le problème est de créer un modèle permettant d'assigner des éléments à ces groupes. Dans ce cas, la tache est appelée classification supervisée.

I.2) Domaines d'application

La modélisation et la classification des données de grande dimension intervient dans de nombreux domaines d'application. Le champ de mise en œuvre de ces techniques va des applications courantes, telles que le traitement des courriers électroniques indésirables **[1]**ou la reconnaissance optique de caractères**[2]**, aux applications professionnelles, telles que l'aide au diagnostic en imagerie médicale**[3]** ou la catégorisation automatique des images d'un satellite**[4].**

I.3) Les étapes d'une classification [5]

1. Choix des données.

2. Calcul des similarités entre les n individus à partir des données initiales.

3. Choix d'un algorithme de classification et exécution.

4. L'interprétation des résultats :

- ✓ évaluation de la qualité de la classification,
- ✓ description des classes obtenues.

I.4)I.4 La classification supervisée

La classification supervisée défini des règles permettant de classer des objets dans des classes à partir de variables qualitatives ou quantitatives caractérisant ces objets.
Il est fortement nécessaire de vérifier la fiabilité de ces dits règles pour les comparer, les appliquer et d'évaluer les cas de sous apprentissage ou de sur apprentissage. On utilise souvent un deuxième échantillon indépendant, dit de validation ou de test.
Plusieurs algorithmes de classification supervisée ont été développés, on cite :

I.4.1) Méthodes probabilistes (statistiques) ou paramétriques

Pour ces méthodes on pose souvent l'hypothèse par exemple gaussienne (moyenne, écart type, variance, covariance).
A fin d'avoir une idée sur le rassemblement des nuages de points, on effectue un examen des éléments statistiques. Pour cela, nous avons divers types de traitements pour pouvoir discriminer ces classes.
Parmi les méthodes probabilistes les plus utilisées on cite :

- ✓ La distance de Mahalanobis,
- ✓ La méthode Parallelepipedique,
- ✓ La méthode du Minimum de distance,
- ✓ La méthode du Maximum de vraisemblance.

I.4.2) Méthodes géométriques ou non paramétriques

Ces méthodes sont souvent sollicitées pour le faible coût en termes de temps de calcul. On peut citer par exemple :

a. Méthode de Sebestien
b. K-plus proches voisins (K-ppv)
c. Méthode barycentrique
d. Méthode Elliptique

- **K-plus proches voisins (K-ppv)** : est une méthode supervisée et non-paramétrique. 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. 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.

  Cette méthode est un moyen simple d'estimation non paramétrique de densité. Pour estimer la densité ri de la classe Ci au point x, on recherche les K (K fixé à l'avance) plus proches voisins de x dans un ensemble de référence (l'ensemble d'apprentissage dont l'affectation des individus est connu à priori). L'estimation de la densité est donnée par :

$$r_i = \frac{K_i\ (x)}{n_i V\ (x)}$$

  où : Ki(x) est le nombre de points de Ci appartenant aux kppv de x,
  ni est le cardinal de la classe Ci,
  V(x) est le volume de la plus petite boule contenant les kppv de x.

Cette méthode se simplifie en méthode de discrimination par voisinage en affectant à x la classe majoritaire parmi les kppv. Le résultat de la discrimination dépend de la valeur de K. C'est pourquoi il est intéressant de faire varier K afin d'obtenir les meilleurs résultats possibles.

Les principaux inco

nvénients de cette méthode sont le coût de stockage (les éléments de l'ensemble d'apprentissage doivent être stockés) ainsi que le coût élevé de la recherche des kppv **[13].**

I.5)Classification non supervisé ou le Clustering

Cette approche déborde le cadre strictement exploratoire. C'est une étude non supervisée, visant à organiser un ensemble d'objets en groupes ou clusters. Ceci est fait en optimisant un critère visant à regrouper les individus dans des classes, chacune le plus homogène possible et, entre elles, les plus distinctes possible.

Le clustering ne nécessite aucun apprentissage et aucune tâche préalable d'étiquetage manuel. Elle consiste à représenter un nuage des points d'un espace quelconque en un ensemble de groupes appelé Cluster.

* Un «Cluster» est une collection d'objets qui sont «similaires» entre eux et qui sont «dissemblables » par rapport aux objets appartenant à d'autres groupes.

I.5.1) Définition du clustering

Le clustering est un processus qui regroupe un ensemble d'objets (physiques ou abstraits) en clusters similaires de telle sorte que les données du même cluster aient des caractéristiques similaires, et celles appartenant à des clusters distincts soient dissimilaires. [26]

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

De plus, les objets dont la classe est inconnue sont généralement disponibles en grand nombre. C'est pourquoi, les approches non supervisées sont massivement utilisées pour traiter des données de manière automatique. C'est la raison pour la quelle nous allons nous intéressés au clustering dans notre mémoire.

I.5.2) Mesures de distance :

Étant donné que le clustering est basé sur le groupement d'objets semblables, la définition d'une mesure qui peut déterminer si deux objets sont semblables ou diffèrent est nécessaire.

Plusieurs méthodes de clustering utilisent les mesures de distance pour statuer sur la similitude ou la dissimilitude de n'importe quelle paire d'objets. La mesure de distance entre deux objets xi et xj est notée d(xi, xj). Une telle mesure devrait être symétrique et atteindre sa valeur minimale (habituellement zéro) en cas de vecteurs identiques. **[13]**

On peut citer

a. La distance de Minkowski :

Etant donnés deux éléments de dimension p, $x_i = (x_{i1}, ...,x_{ip})$ et $x_j= (x_{j1},...,x_{jp})$, la distance entre les deux éléments peut être calculée par la métrique de Minkowsk :

$$d(xi, xj) = |x_{i1} - x_{j1}|^g + |x_{i2} - x_{j2}|^g + ... + |x_{ip} - x_{jp}|^g$$

- ✓ Pour g=2, on obtient la distance Euclidienne.
- ✓ Pour g=1, la mesure est appelée distance de Manhattan.
- ✓ Pour g = ∞, c'est la métrique de Tchebychev

I.5.3) Les méthodes de clustering

Les méthodes se limitent à l'exécution d'un algorithme itératif convergeant vers une bonne partition et correspondant en général à un optimum local.

Plusieurs choix sont laissés à l'initiative de l'utilisateur :

- ✓ une mesure d'éloignement (dissemblance, dissimilarités ou distance) entre individus;
- ✓ le critère d'homogénéité des classes à optimiser : il est, dans le cas de variables quantitatives, généralement défini à partir de la trace d'une matrice de variances-covariances; soit les variances et covariances interclasses (la trace correspond alors à l'inertie de la partition), soit les variances et covariances intra classe;
- ✓ la méthode : classification ascendante hiérarchique, réallocation dynamique et DBSCAN sont les plus utilisées, seules ou combinées;
- ✓ le nombre de classes : c'est un point délicat.

Les méthodes de la classification non supervisée ou du clustering sont généralement classifiées en quatre catégories majeures :

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

I.5.4) Les méthodes hiérarchiques.

Dans un clustering hiérarchique, un cluster peut être divisé en sous clusters, l'ensemble des clusters étant généralement représenté par un arbre. Un objet appartient a une et une seule feuille dans la hiérarchie, mais également a son nœud père, et ainsi de suite jusqu'a la racine.

Il existe deux types d'approches de clustering hiérarchique :

a. Les approches par agglomération (Bottom up approach) :

La classification ascendante hiérarchique (CAH) est une méthode de classification itérative dont le principe est simple, il commence par des clusters formes d'un seul objet, puis les fusionne successivement jusqu'a ce que le critère d'arrêt soit atteint (Construction de K clusters par exemple). Les méthodes d'agrégation pour la CAH sont :

- Méthode de Ward
- Lien fort
- Lien faible
- Lien du centre de gravité
- Lien moyen

La classification ascendante hiérarchique (CAH) présente

Les ***avantages suivants :***

- Pas besoin de fixé le k en avance.
- Meilleur analyse et compréhension
- On travaille à partir des dissimilarités entre les objets que l'on veut regrouper. On peut donc choisir un type de dissimilarités adapté au sujet étudié et à la nature des données.
- L'un des résultats est le dendrogramme, qui permet de visualiser le regroupement progressif des données. On peut alors se faire une idée d'un nombre adéquat de classes dans lesquelles les données peuvent être regroupées [11].

Et les Inconvénient suivants:

- Calcul couteux
- On ne peut pas défaire les fusions.

b. Les approches par division (Top down appoach) :

Cette approche commence par un cluster forme de tous les objets, qui sera ensuite divise en petits clusters jusqu'a atteindre une condition d'arrêt donnée par l'utilisateur **[8].**

Les avantages :

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

- La lecture de l'arbre permet de déterminer le nombre optimal de classes.

Les inconvénients :

- Imprécision sur les critères d'arrêt.
- Coûteux en temps de calcul.
- La plupart des algorithmes hiérarchique ne revisitent pas les clusters une fois construits

I.5.5) Les méthodes de partitionnement

L'ensemble des M clusters résultants à l'application de ces méthodes sont représentés par des centroides (représentant les clusters), chaque objet appartenant à un seul est un seul cluster. Un centroide est considéré comme une description des caractéristiques commune de tous les objets du même cluster. Ces caractéristiques correspondent le plus souvent à des critères de proximité (similarité informatique) que l'on définit en introduisant des mesures et classes de distance entre objets**. [8]**

Pour obtenir un bon partitionnement, il convient à la fois de :

✓ Minimiser l'inertie intra-classe pour obtenir des grappes (cluster en anglais) les plus homogènes possibles ;

✓ Maximiser l'inertie inter-classe afin d'obtenir des sous-ensembles bien différenciés.

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

a. **Méthode du K-Means (Macqueen, 1967) :** L'algorithme du k-means est le plus populaire des algorithmes de clustering, il partitionne un ensemble de données en un nombre prédéfini de régions K. Chaque groupe est représenté par sa moyenne (centre de la classe) qui ses coordonnées sont la moyenne arithmétique pour chaque dimension séparément de tous les points dans le cluster.

   Dans cette méthode, un cluster est représenté par son centroide qui est une moyenne (habituellement pondérée) des points situées a l'intérieur du cluster, cette approche ne fonctionne convenablement qu'avec les attributs numérique et le résultat final peut être négativement affecté par la présence de bruits.

   La somme des écarts entre un point et son centroide, exprimée avec une mesure appropriée, est utilisée comme fonction objectif. Chaque point est assigné au cluster dont le centroide est le plus proche**. [8]**

Algorithme : K-means ----------

1. phase d'initialisation :

- Générer K points comme centroides ou référent au hasard.

2. phase d'affectation :

- Calculer la distance entre les points et chaque centroide.
- Affecter le point au cluster dont la distance à son cenroide est la plus petit.

3. phase de minimisation :

Pour chaque cluster calculer la moyenne de tous ces point pour déterminer le nouveau centroide.

4. Répéter les étapes 2 et 3 jusqu'à ce qu'aucun affectation au clusters ne change plus.

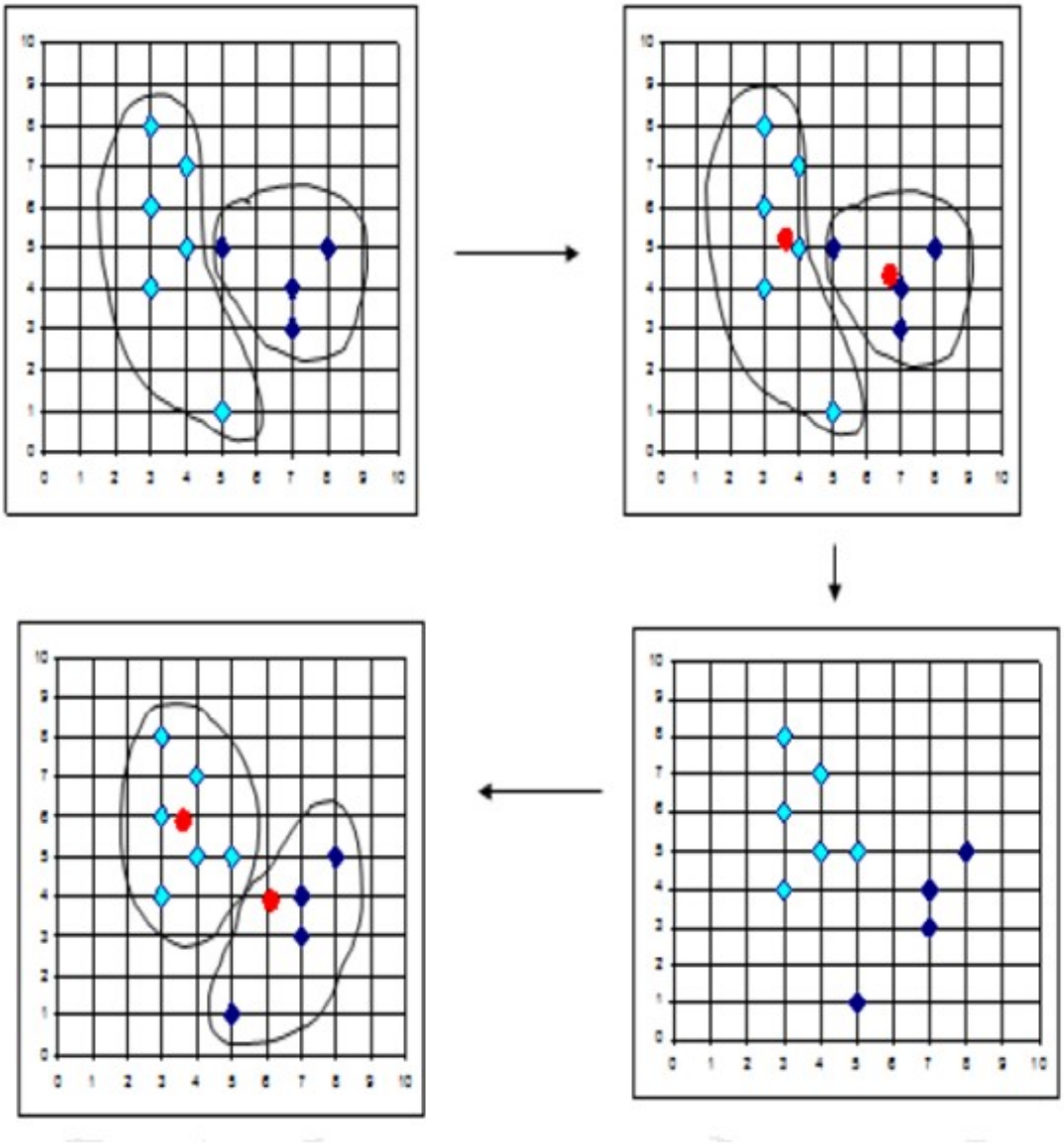

Figure I) 1- Exemple sur l'algorithme K-means

La classification k-means présente notamment les avantages **[8][12]** suivants :

- ✓ Un objet peut être affecté à une classe au cours d'une itération puis changer de classe à l'itération suivante, ce qui n'est pas possible avec la classification ascendante hiérarchique pour laquelle une affectation est irréversible.
- ✓ En multipliant les points de départ et les répétitions on peut explorer plusieurs solutions possibles.
- ✓ Facile à implémenter ;
- ✓ Fonctionne avec toutes les mesures standards ;
- ✓ Insensible à l'ordre des données.
- ✓ Calcul rapide.
- ✓ On peut défaire les fusions.

L'inconvénient de cette méthode **[12][8]** est

- ✓ qu'elle ne permet pas de découvrir quel peut être un nombre cohérent de classes, ni de visualiser la proximité entre les classes ou les objets. Les méthodes kmeans et la CAH sont donc complémentaires.
- ✓ Il n'est pas applicable en présence d'attributs qui ne sont pas du type numérique
- ✓ On doit fixer k en avance.
- ✓ Le résultat final dépend fortement du choix des centroides initiaux
- ✓ Sensible à la présence de bruits

- On propose une autre version de k means inspiré du principe de vote majoritaire **[19]** qui sera appliqué dans notre travail avec le k meams classique, les résultats seront présentés dans le troisième chapitre.

Algorithme vote ----------

Structure :

PopulationElement: List<Float> attributs ;

Population: List<PopulationElement> ;

Observation : List<Integer> ;

Simulations : List<Observation> ;

Entree: population D ;

Nombre de cluster K ;

Variable List<Observation> ob ;

Debut

D = lirePopulation(fichierCSV) ;

K = lireK() ;

Nombre_attribut = detecter_nombre_attribut(d.lirePremierelement()) ;

Pour i allant de 0 a nombre_attribut

Pour j allant de i a nombre_attribut -1 faire

Population_pair_I_J = former_population_depuis(i,j);

Ob = Appliquer_kmeans(Population_pair_I_J) ;

Simulation.ajouter(ob) ;

Fin pour

Fin pour

Classification_final = Simulation.appliquer_vote() ;

Fonction detecter_nombre_attribut(List<Float> attributs)

Debut

Retourner attributs.taille ;

Fin

I.5.6) Méthodes basée sur la densité

Ces méthodes découvrent les clusters de formes arbitraires et assure l'isolement des bruits.

La fonction de densité est définie comme nombre d'objet dans un voisinage spécifique des éléments de donnés. Elle est basée sur le rapport entre le nombre de point présents dans le cluster et son volume.

Ces algorithmes basés sur la densité utilisent cette fonction de densité pour grouper les objets dans les clusters. [**14]**

Dans cette approche, un cluster donné continue à augmenter de taille tant que le nombre d'objets dans le voisinage dépasse un certain seuil.

Elle se subdivise en deux types :

a. **Méthodes basées sur la densité connective (Density-Based Connectivité Clustering) :** la densité et la connectivité sont mesurés en termes de distribution locale des voisins les plus proches.

   Tous les points accessibles à partir des noyaux des objets sont factorisés dans les composants connectés servant de clusters. Les autres points ne sont couverts par aucuns clusters. Ces sont considères comme des bruits (outliers). Les points non fondamentaux à l'intérieur d'un cluster représentent sa borne et les objets du noyau sont les points internes**[14].**

b. **Méthodes basées sur les fonctions densités (DENSITY FUNCTIONS CLUSTERING):** La densité globale est la somme des fonctions de densité des objets. Les clusters sont détermines par les attracteurs de densité qui sont définis comme les maxima locaux de la fonction de densité globale. **[14]**

Example:

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) **[7]**

DBSCAN permet l'identification des clusters de formes arbitraires et le bruit dans une base de données spatiale. Cet algorithme requiert seulement deux paramètres d'entrée afin que l'utilisateur puisse spécifier une valeur appropriée. On fixe Eps, le rayon du voisinage à étudier et MinPts, le nombre minimum de points qui doivent être contenus dans le voisinage pour considérer la zone comme dense. L'idée clé du clustering basé sur la densité est que pour chaque point d'un cluster, ses voisins pour un rayon donné Eps doit contenir un nombre

minimum de points MinP ts. Ainsi, le cardinal de son voisinage doit dépasser un certain seuil (consid_er_e comme objet principal). Ensuite, DBSCAN collecte itérativement et de proche en proche les objets atteignables par densité par rapport aux objets principaux, le processus se termine lorsqu'aucun nouveau point ne peut être ajouté à un cluster. **[8]**

I.5.7) Méthode basée sur les grilles

Un algorithme de clustering basé sur les grilles est fondé sur le principe de la discrétisation de l'espace des données **[6].** L'espace d'objets est quantifié en un ensemble de cellules, puis identifie l'ensemble de cellules denses connectées pour former des clusters.

I.6)Techniques de validation du clustering

L'objectif principal de la validation des clusters est d'évaluer les résultats du processus de clustering afin de choisir le meilleur partitionnement des données **[12].** Par conséquent, des approches de validation sont utilisées pour évaluer quantitativement le résultat d'un algorithme de clustering. Ces approches possèdent des indices représentatifs, appelés indices de validité.

Il existe plusieurs indices de validité, parmi eux on cite :

I.6.1) L'indice de Dunn

L'indice de Dunn tient compte à la fois de la compacité et de la séparabilité des groupes: la valeur de cet indice est plus élevée que les groupes sont compacts et bien séparés.
L'indice de Dunn prend la forme suivante :

$$\text{Dunn}=\frac{Dmin}{D\ max}$$

Où d_{min} est la plus petite distance intraclasse et d_{max} la plus grande distance interclasse
La valeur de cet indice est plus élevée que les groupes sont compacts et bien séparés.

Exemple :
la distance inter _classes δ
la distance intra_ classe Δ

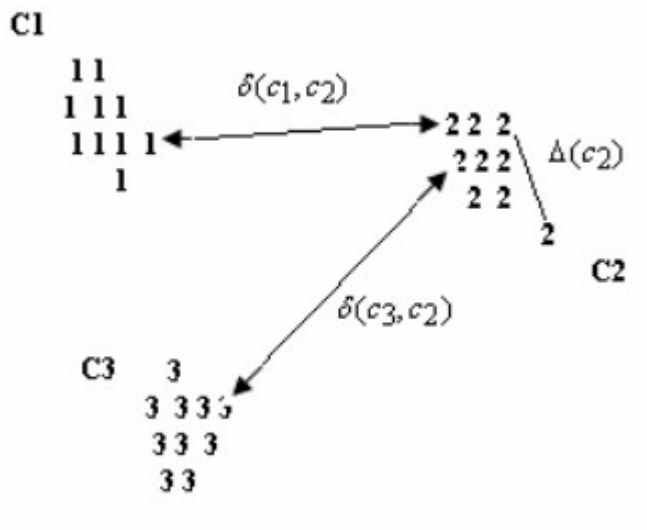

Figure I) 2-l'indice de Dunn

L'objectif principal de l'indice de Dunn est de maximiser les distances inter-clusters (séparation) et de minimiser les distances intra-clusters (augmenter la compacité)

Mais cet indice présente deux inconvénients majeurs :

- ✓ Son calcul est coûteux.
- ✓ Il est sensible `a la présence de bruit.

I.6.2) L'indice C_index

Cet index est défini comme suit:

$$C = \frac{S - S_{min}}{S_{max} - S_{min}}$$

Où

S est la somme des distances, sur toutes les paires d'objet de la même classe.

Soit *L* le nombre de ces couples,S_{min} est la somme des *L* distances les plus petites, S_{min}est la somme des *L* distances les plus larges de tous les couples.

Une faible valeur de C indique un bon regroupement.

I.6.3) Indice de Davies-Bouldin

L'indice de Davies-Bouldin tient compte à la fois de la compacité et de la séparabilité des groupes. Leur valeur est d'autant plus faible que les groupes sont compacts et bien séparés. L'objectif de cet indice est de minimiser la similarité moyenne entre chaque cluster et le cluster qui lui est le plus similaire.

$$DB = \sum_{k=1}^{c} max_{(k \neq l)} \left\{ \frac{\mathrm{sc(k)} + \mathrm{sc(l)}}{\mathrm{dce(k\,,l)}} \right\}$$

Où $s_{c.}$ (c) est la distance moyenne entre un objet du groupe X_i et son centre C_k.

$$S_c = \frac{\sum_i \|x_i - c_k\|}{N_k}$$

N_k : le nombre des éléments dans la classe k

d_{ce} : la distance qui sépare les centres des groupes c_k et c_l

$$d_{ce} = \|c_k - c_l\|$$

Une valeur optimale de K est celle qui minimise DB.

I.6.4) Indice WB

C'est l'indice des sommes carrées (WB), il est défini comme un ratio de la mesure de la compacité du cluster à son mesure de la séparation. L'indice est donné par:

$$\mathbf{SSW} = \sum_{i=1}^{n} \|\boldsymbol{Xi} - \boldsymbol{Cpi}\|^2$$

$$\mathbf{SSB} = \sum_{i=1}^{k} \boldsymbol{Ni} \|\boldsymbol{Ci} - \bar{\boldsymbol{x}}\|$$

$$\mathbf{WB} = \mathbf{K} \frac{\boldsymbol{SSW}}{\boldsymbol{SSB}}$$

Où

K : représente le nombre de clusters.

N : le nombre total de points.

$n_{i\,:}$ représente le nombre de points dans le cluster c_i

X : barre c'est la moyenne de l'ensemble de données X

C_i : est le centre du cluster i

C_{pi} : est le centre du cluster i

Une faible valeur de l'indice WB indique un bon regroupement.

I.7) Conclusion

Nous avons présenté dans ce chapitre un ensemble de méthodes de classification supervisé et non supervisé (clustering). Ces méthodes se basent sur des techniques très différentes et donnent des résultats dont la qualité dépend des paramétrages d'initialisation choisis.

Ensuite, nous avons prolongé notre attention sur les principales méthodes du clustering.

Il en ressort qu'aucune méthode ne peut être qualifiée de meilleurs par rapport à une autre. Pour tenter de résoudre ce problème, nous allons voir dans le chapitre suivant comment introduire l'approche de la théorie des jeux dans le clustering. La mise en œuvre de cette approche en clustering nous donne un espoir d'améliorer ce dernier pour produire des résultats plus pertinents.

II. La théorie des jeux

II.1) Introduction :

En 1944, Von Neumann et Oskar Morgenstern postulent dans un ouvrage fondateur (Theory of Games and Economic Behavior, Princeton University Press) que tous les problèmes économiques peuvent se rapporter à un jeu de stratégie entre des acteurs rationnels (joueurs). La théorie des jeux devient le moyen d'analyser les interactions entre les joueurs (une révolution des sciences économiques).

La théorie des jeux est un outil d'analyse et de modélisation des comportements rationnels de joueurs en situation d'interaction stratégique. Elle s'intéresse à des situations où des joueurs ou « agents » prennent des décisions, chacun étant conscient que ses gains dépendent non seulement de sa propre décision, mais aussi des décisions prises par les autres joueurs. Un joueur peut prendre plusieurs décisions et il en choisit une qui lui semble la meilleure, appelée stratégie. En termes mathématiques, on traduit la phrase « la meilleure décision pour lui » par l'introduction d'une fonction pour chacun des joueurs qui reflète ses préférences, appelée fonction de gain ou fonction d'utilité.

Vers les années 50, John Nash prouvait que tout jeu possède une situation d'équilibre mixte, dite d'équilibre de Nash, dans lequel aucun joueur n'a intérêt à s'écarter unilatéralement.

II.2) A quoi sert la théorie des jeux ?

. Jeux de société (échecs, dames, go, ...), Jeux de cartes (bridge, poker, ...)

. Dois-je travailler ou faire semblant ?

. Est-ce que j'écoute de la musique ce soir ?

. Enchères, vote

. Comportement animal

. Stratégies militaires/économiques

. Partages de ressource (marchandage)

. Est-ce qu'une entreprise doit exploiter ses salariés ?

. Est-ce qu'une entreprise doit entrer sur un marché ou pas ?

. Faut-il contrôler les déclarations d'impots sur le revenu ?

II.3) Définition d'un jeu

Un jeu est définit comme un univers dans lequel chaque preneur de décision possède un ensemble d'actions possibles déterminé par les règles du jeu. Le résultat du jeu dépend alors conjointement des actions prises par chaque preneur de décision.

Donc un jeu est défini par l'ensemble de joueurs, leur ensemble de stratégies et la fonction de gains. Il est modélisé par le triplet : $G = (N, (S_i)_{i \in N}, (u_i)_{i \in N})$

Où : $N = \{1,\ldots, n\}$ est l'ensemble de joueurs

S_i est l'ensemble des stratégies possibles du joueur i tel que $i \in N$

s_i désigne une stratégie du joueur i tel que $i \in N$ et $s_i \in S_i$

$s = (s_1,.., s_n) \in S1 \times \ldots \times Sn$: est une issue de jeu (une situation possible)

$u_i(s) \in R$ est la fonction de gain du joueur i tel que $i \in N$

II.4) Fonction de gains (fonction d'utilités)

La satisfaction d'un joueur peut être représentée par un nombre positif ou négatif. Plus le nombre est élevé, plus le joueur est satisfait. Ces gains sont définis par une fonction de gains**.[20]**

Pour définir une fonction de gain, il faut prendre en compte les facteurs suivants :

- Rationalité des joueurs. Ils veulent arriver à une issue du jeu qui soit la meilleure pour eux.
- Nécessité de connaître les gains de chaque joueur sur les issues possibles du jeu.
- Mesure de la satisfaction d'un agent face à chaque issue du jeu.

La fonction de gain est utilisée pour choisir quelle stratégie doit adopter chacun des joueurs (maximiser son gain). Elle lie un ordre de gains à des valeurs numériques.

II.5) La modélisation des jeux

On peut diviser les représentations des jeux en deux classes:

II.5.1) Forme stratégique

Un jeu sous forme stratégique (ou jeu sous forme normale) est défini par :

- Un ensemble de joueurs,
- l'ensemble des stratégies possibles pour chacun des joueurs,

- les préférences de chacun des joueurs sur l'ensemble des combinaisons stratégiques possibles.

Le modèle de jeux sous forme stratégique oblige à supposer que les joueurs choisissent leur stratégie une fois pour toute. Ils sont alors engagés dans cette stratégie et ne peuvent pas la modifier à mesure que le jeu se déroule.

Cette modélisation est toutefois très utile pour décrire des situations dans lesquelles les joueurs jouent en même temps. Le plus connu des jeux sous forme stratégique est le dilemme du prisonnier. Ce jeu, joué entre deux suspects d'un crime, tire son importance du fait qu'il existe un très grand nombre de situations dans lesquelles le même type d'incitations est présent

Exemple : Le dilemme du prisonnier : Deux suspects d'un crime majeur sont détenus dans des cellules séparées. La police a assez de preuves pour condamner chacun d'entre eux pour des crimes mineurs mais pas assez pour les condamner pour le crime majeur, à
moins que l'un d'entre eux ne dénonce l'autre. Si les deux suspects se taisent, ils seront chacun condamnés à un an de prison. Si seulement l'un d'entre eux dénonce l'autre, il sera libéré et utilisé en tant que témoin contre l'autre qui écopera de 10 ans de prison. Enfin si les deux dénoncent, ils passeront chacun 5 ans en prison.
Ce jeu peut être représenté comme un jeu stratégique où :

- les joueurs sont les deux suspects
- qui ont chacun le choix entre deux actions : {Se taire, Dénoncer}
- on suppose que les préférences des joueurs sont uniquement déterminées par les années qu'ils passeront en prison. Ainsi :

u1 (Dénoncer, Se taire) > u1 (Se taire, Se taire) > u1 (Dénoncer, Dénoncer) > u1(Se taire, Dénoncer), et
u2 (Se taire, Dénoncer) > u2 (Se taire, Se taire) > u2 (Dénoncer, Dénoncer) > u2(Dénoncer, Se taire)

SUSPECT 2

SUSPECT 1	Se taire	Dénoncer
Se taire	(-1,1)	(-10,0)
Dénoncer	(0,-10)	(-5,-5)

II. 1 tableau du lemme du prisonnier

Il est usuel de représenter un jeu à deux joueur sous forme stratégique par le tableau des gains. Le jeu peut alors être représenté comme suit :

II.5.2) Forme extensive

Lorsque les règles du jeu stipulent que les joueurs interviennent les uns après les autres, dans un ordre précis et que le nombre d'actions parmi lesquelles leur choix s'exerce est fini, la représentation qui semble la plus appropriée consiste à tracer un arbre. Une telle représentation, dite sous forme extensive. **[20]**

Représentation graphique : Tout jeu sous forme extensive peut être représenté par un arbre (graphe connexe sans cycle) où

- à chaque nœud terminal correspond un résultat du jeu
- à chaque nœud non terminal est associé un joueur : arrivé à ce point du jeu, c'est à son tour de jouer.
- chaque arc représente chacune des actions que ce joueur peut prendre à ce point du jeu

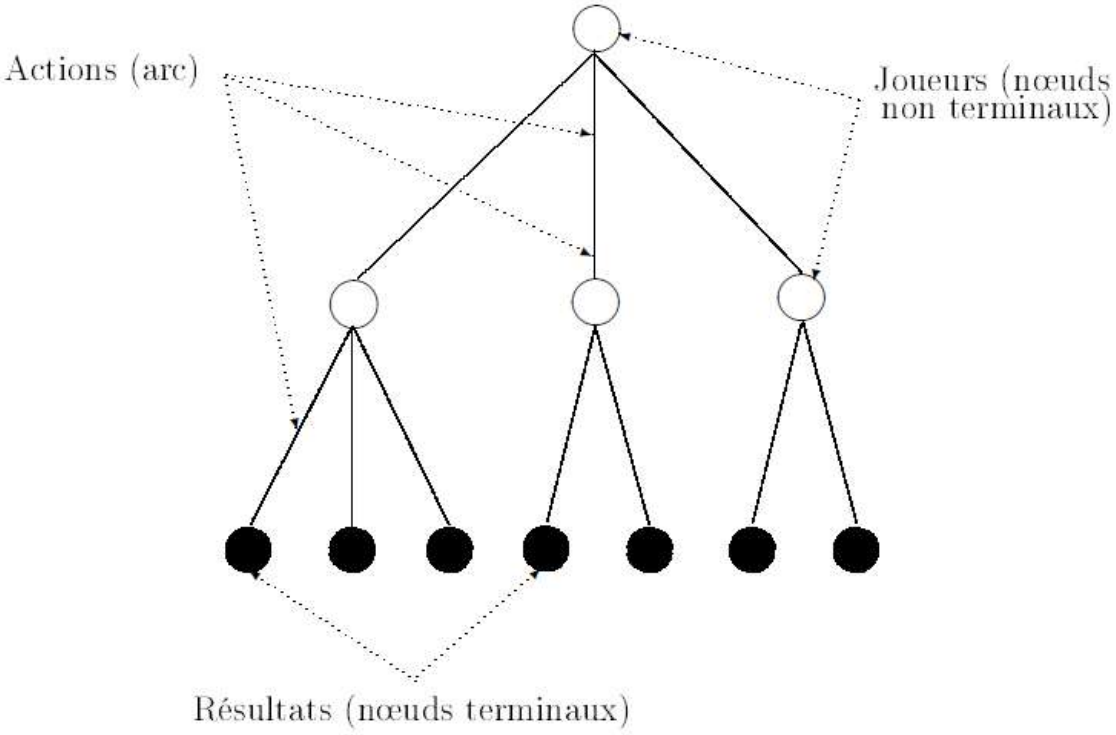

Figure II) 1-jeux forme extensive

II.6) Typologies des jeux

II.6.1) Jeux coopératifs / non coopératifs

Les jeux coopératifs reposent sur la poursuite d'un objectif commun pour tous les joueurs. Cet objectif ne pourra être réalisé que par l'entraide et la solidarité entre eux. Il ne s'agit pas de gagner sur l'adversaire mais de faire équipe pour gagner ensemble... ou de perdre ensemble si l'équipe s'est mal organisée **[21].**

Les jeux non coopératifs modélisent les interactions où les joueurs sont libres de choisir leurs actions et où un joueur rationnel cherche à maximiser son propre bien- être (si un joueur se rend compte qu'il a une stratégie admissible b lui permettant d'obtenir une meilleure utilité que celle obtenue avec la stratégie a alors il ne devrait pas jouer a**)[21].**

II.6.2) Jeux avec décisions simultanées / séquentielles

Dans un jeu, si les joueurs décident de leurs actions simultanément, alors on parle dans ce cas de *jeu simultané*. A l'inverse, si les joueurs décident de leur actions l'une aprés l'autre, alors on est dans le cas *d'un jeu séquentiel.*

II.6.3) jeux à information complète / incomplète

Un jeu est dit à *information complète* si chacun des joueurs connaît la structure du jeu. Le jeu du dilemme du prisonnier est à information compète car chacun des prisonniers connait parfaitement la régle du jeu définie par le policier ainsi que l'utilité de l'autre joueur.

Si au moins un des joueurs ne connaît pas entièrement la structure du jeu, le jeu est dit à *information incomplète.*

II.6.4) Jeux à information parfaite / imparfaite

Un jeu est dit à *information parfaite* si chacun des joueurs, au moment de choisir son action, a une connaissance parfaite de l'ensemble des décisions prises antérieurement par les autres joueurs, comme exemple la dame, jeu d'échec …
Un jeu est à information imparfaite si un des joueurs ne connaît pas, ce qu'a joué un autre joueur. Ceci peut arriver dans le cas où on cache l'information aux joueurs ou parce que les joueurs jouent simultanément. Le jeu du dilemme du prisonnier est à information imparfaite car les deux joueurs jouent simultanément.

II.6.5) Jeux répétés

Les jeux répétés sont des jeux qui sont joués plus d'une fois. Les joueurs peuvent choisir des actions différentes en considérant l'histoire du jeu étant donné que l'expérience que les joueurs ont acquise à travers la répétition est cruciale pour définir leurs actions futures.

II.7) Equilibre de Nash (Un homme d'exception)

Du remarquable livre de Sylvia Nasar est tiré le non moins remarquable film hollywoodien : A beautiful mind (Un homme d'exception) qui relate la vie du plus célèbre théoricien des jeux : nous avons nommé John Nash (interprété à l'écran par Russel Crow). John Nash est l'étudiant torturé et schizophrène de John Von Neumann. Il sera interné pendant plus de 20 ans après avoir péniblement publié deux papiers et une thèse en 23 pages… de qualité ! Il sortira de l'hôpital prix Nobel, pour avoir notamment trouvé une façon « équilibrée » (contrastant avec sa personnalité) de résoudre n'importe quel jeu et n'importe quel dilemme : « tout jeu admet au moins un équilibre de Nash ».

L'équilibre non coopératif, dit aussi équilibre de Nash, est basé sur le principe de rationalité individuelle. Il s'agit d'un état dans lequel aucun joueur ne souhaite modifier sa stratégie si les autres joueurs maintiennent leurs stratégies d'équilibre. Chaque stratégie est une meilleure réponse aux stratégies des autres joueurs.

II.8) Le clustering par la théorie des jeux

Ces dernières années, la théorie des jeux avec ses différentes branches, a constitué une boite à outil servant de support au traitement du problème de clustering. Plusieurs auteurs (Swapnil Dhamal **[22]**, Satyanath Bhat **[23]** , Anoop K R **[24]**, et Varun R Embar)[**25],** se sont intéressés à cette approche.

II.8.1) Kmeans et la théorie des jeux

Dans cette partie, nous présentons une nouvelle approche, qui modélise le problème de clustering dans un contexte de jeu. Puis, nous évaluons son efficacité par rapport au kmeans classique.

Notre approche inspirée des travaux **[8][27]** consiste à trouver une partition de l'espace de départ telle que les données appartenant à un même groupe soient plus similaires entre elles qu'avec les données issues d'un autre groupe. Elle nous permet de construire K partitions (clusters) initiales d'individus similaires et les améliorer afin d'obtenir des clusters correspondant à un équilibre du jeu associé.

II.8.2) Modélisation de problème sous forme d'un jeu

Notation

- Description des clusters :

Considérons un nombre K de clusters, qui sont dénotés par C = {C1; C2; ….; CK} . Soit D = {1; 2; …. ; n} un ensemble d'objets de cardinalité n. notons n >> K. Chaque joueur i est caractérisé par des attributs $\{t_{i;1}$; $t_{i;2}$; …. $t_{i;\delta}$ }, où δ est la dimension du vecteur caractéristique. Ces attributs seront utilisés pour définir la mesure de voisinage d'un objet parmi les données.

Soit P = {cj / j = 1; …K}, notons que cj est le représentant du cluster Cj . La distance entre deux équipes est exprimée par la distance qui sépare leurs représentants respectifs.

- Les stratégies :

Le jeu est en forme stratégique, on pose une collection de stratégies décrivant les actions de chaque joueur dans toutes les situations concevables du jeu, ainsi que les gains que chacun obtient lorsque les stratégies de tous les joueurs sont connues. Seules certaines actions seront effectivement choisies. Les stratégies sont les suivantes :

1. La stratégie " Vendre un joueur " notée par x1 : cela signifie qu'il y a deux équipes dont l'une d'elle accepte de vendre un joueur et l'autre de l'acheter ;

2. La stratégie " Acheter un joueur " notée par x2 : cela signifie qu'il y a au moins deux équipes dont l'une d'elle accepte d'acheter un joueur et l'autre de le vendre ;

3. La stratégie " Virer un joueur " notée par x3 : cela signifie que le joueur est viré de sa propre équipe et ne peut pas être admis par aucune des autres équipes, car l'ajout de cet objet (joueur) n'améliore pas leur homogénéité. Ainsi, l'objet est mis dans une équipe à ajouter (la K_{me+1} équipe) portant l'étiquette 'Corbeille' ;

4. La stratégie " Inscrire un joueur " notée par x4 : cela signifie qu'une équipe a décidé de prendre l'un des joueurs de l'équipe 'Corbeille', car il améliore son homogénéité ;

5. La stratégie " Ne rien faire " notée par x5 : l'équipe est satisfaite des joueurs dont elle dispose.
Si on note par Xi l'ensemble des stratégies du joueur 'i', alors X1 = X2 = … = Xn = {x1; x2; x3; x4; x5}

Le premier objectif consiste à construire une équipe homogène, dont ses membres ont tendance à avoir une forte similarité entre eux, en d'autre terme, le joueur participant tente de jouer aussi longtemps que possible, en choisissant l'une des stratégies qui améliore son homogénéité, dans le cas où il ne joue pas, la stratégie correspondante est "ne rien faire".

II.8.3) Algorithme notre approche

Dans ce qui suit, nous allons présenter l'algorithme k-means à la base de la théorie des jeux :

Entrée : ***D*** *:* Base de donnée (les joueurs)

MaxIter : le nombre d'itérations

SeuilMin : le nombre minimum de joueurs dans une équipe (=2)

Sortie : l'ensemble des K équipes

Pour K = 2 jusqu'à 10

K : le nombre d'équipes (clusters)

1. Choisir **k** joueurs représentant alétoires
2. Calculer les voisins de chaque joueur choisis notée : ***Vois(i)***
3. Calculer le représentant (le centre de gravité ou la moyenne) de chaque équipe (cluster)
4. Affecter les joueurs aux équipes les plus proches
5. **Répéter**
   - Calculer les capitaux de chaque objet joueur pour chaque cluster

$$\lambda\ (i, Cj) = \frac{1}{|\mathrm{Cj}|} \sum_{l \in Cj} d(i, l)$$

   |Cj| la taille d'un cluster Cj (équipe)
   d(i, l) est la distance euclidienne entre deux objets i et l
   La mesure qui est largement déployée pour le calcul de la similarité est la distance euclidienne

   - Pour chaque équipe Cj
     Choisir une des stratégies **Vendre (), Acheter (), Virer () ou Inscrire ()**
   - Calculer l'indice de validité (le gain)
6. **Jusqu'à atteindre** le nombre d'itérations maximal ***MaxIter*** ou choisir l'itération dans laquelle l'indice est optimal

Prendre la classification de K ou l'indice est optimal

Fin de pour

II.9) La différence entre le supervisé et non supervisé

La différence entre les deux situations est la connaissance des classes

- ✓ Dans le cas non supervisé, on fait une recherche à l'aveugle alors en qu'en supervisé on a un échantillon d'apprentissage
- ✓ En supervisé on connait un indicateur de performance : le taux de mal classé, mais en non supervisé ? Difficile de valider les résultats **[29]** .

II.10) Conclusion

L'intérêt principal de la théorie des jeux est d'étudier les différentes situations d'interactions stratégiques entre des individus, où chaque individu cherche à maximiser son gain personnel. Nous avons constaté que l'analyse des jeux est basée sur la notion d'équilibre, et en particulier l'équilibre de Nash qui permet de définir une situation de non regret pour les différents joueurs.

Cependant, l'équilibre trouvé n'assure pas le gain maximal pour tous les joueurs. Plusieurs équilibres peuvent être trouvés pour un jeu. Par conséquent, la question qui se pose est la suivante *« quel est le meilleur équilibre ? ».*

Dans ce chapitre, nous avons proposé une technique de clustering, nous avons exploité une technique standard, à savoir, le k-means à laquelle nous avons introduit une nouvelle approche de recherche qui se base sur la théorie des jeux.
L'expérimentation de cet algorithme et les résultats seront présentés dans le chapitre suivant.

III. Application

III.1) Introduction

Ce chapitre présente les outils de développement et les différents algorithmes que nous avons utilisés dans notre application.

III.2) Langage et outils de développement

Parmi les différents outils de développement, nous avons choisi ces outils que nous avons utilisés pour réaliser notre travail.

a- Netbeans : L'EDI NetBeans est un environnement de développement intégré, gratuit à l'usage, se concentrant principalement sur simplifier le développement d'applications Java. Il fournit du support pour tous les types d'applications Java, depuis le client riche jusqu'aux applications d'entreprises multicouches, en passant par les applications pour les mobiles supportant Java.

L'EDI NetBeans a une architecture modulaire qui permet les 'plug-ins'.

L'EDI NetBeans fournit un étalage étonnant de fonctionnalités directement prêtes à l'usage (out of the box). L'EDI NetBeans possède un environnement de développement entièrement dédié à J2EE. Tout le support de projet, éditeur, débogage, qui sont disponible pour le développement d'applications Java est également disponible pour le développement J2EE. De plus, l'EDI NetBeans fournit l'accès au Catalogue de solutions Java BluePrints et la capacité de les installer en tant que projet NetBeans**. [15]**

NetBeans contient, en plus du support pour CVS et SubVersion, un support pour ClearCase, mais aussi pour Mercurial.

Enfin cet IDE possède un débogueur de grande qualité ainsi qu'une interface graphique améliorée.

b- **Maven**

Maven est un outil de construction de projets (build) open source développé par la fondation Apache, initialement pour les besoins du projet Jakarta Turbine.

Il permet notamment :

- d'automatiser certaines tâches : compilation, tests unitaires et déploiement des applications qui composent le projet
- de gérer des dépendances vis-à-vis des bibliothèques nécessaires au projet
- de générer des documentations concernant le projet

Au premier abord, il est facile de croire que Maven fait double emploi avec Ant. Ant et Maven sont tous les deux développés par le groupe Jakarta, ce qui prouve bien que leur utilité n'est pas aussi identique que cela. Ant, dont le but est d'automatiser certaines tâches répétitives, est plus ancien que Maven. Maven propose non seulement les fonctionnalités d'Ant mais en propose de nombreuses autres.

-Pour gérer les dépendances du projet vis-à-vis de bibliothèques, Maven utilise un ou plusieurs dépôts qui peuvent être locaux ou distants.

-Maven est extensible grâce à un mécanisme de plugins qui permettent d'ajouter des fonctionnalités.

Maven utilise une approche déclarative où la structure du projet et son contenu sont décrits dans un document XML. De plus il convient de se conformer à une structure de projets standards et de bonnes pratiques. L'observation de ces normes permet de réduire le temps nécessaire pour écrire et maintenir les scripts de build car ils sont tous structurés de la même façon.

Maven peut aussi assurer de nombreuses autres tâches car il est conçu pour utiliser des plugins : il est donc extensible. Maven est fourni avec un grand nombre de plugins standard mais il est aussi possible d'utiliser d'autres plugins qui sont stockés dans les dépôts voire même de développer ses propres plugins**.[16]**

c-GIT : Un logiciel de gestion de versions pour gérer le code source d'OpenClassrooms.

Type de GIT : distribué.

Description de GIT :très puissant et récent, il a été créé par Linus Torvalds,qui est entre autres l'homme à l'origine de Linux. Il se distingue par sa rapidité et sa gestion des branches qui permettent de développer en parallèle de nouvelles fonctionnalités.

Avantage de Git :

- est très rapide ;
- sait travailler par branches (versions parallèles d'un même projet) de façon très flexible ;
- est assez complexe, il faut un certain temps d'adaptation pour bien le comprendre et le manipuler, mais c'est également valable pour les autres outils ;
- est à l'origine prévu pour Linux. Il existe des versions pour Windows mais pas vraiment d'interface graphique simplifiée. Il est donc à réserver aux développeurs ayant un minimum d'expérience et… travaillant de préférence sous Linux**.[17]**

Une des particularités de Git, c'est l'existence de sites web collaboratifs basés sur Git comme GitHub. GitHub, par exemple, est très connu et utilisé par de nombreux projets : jQuery, Symfony, Ruby.

-C’est une sorte de réseau social pour développeurs : vous pouvez regarder tous les projets évoluer et décider de participer à l’un d’entre eux si cela vous intéresse.

-Vous pouvez aussi y créer votre propre projet : c’est gratuit pour les projets *open source* et il existe une version payante pour ceux qui l’utilisent pour des projets propriétaires.

-Si deux personnes modifient en même temps deux endroits distincts d’un même fichier, les changements sont intelligemment fusionnés par Git.

d-JavaFX : JavaFX est une technologie créée par Sun Microsystems qui appartient désormais à Oracle. Avec l'apparition de Java 8 en mars 2014, JavaFX devient la bibliothèque de création d'interface graphique officielle du langage Java, pour toutes les sortes d'application (applications mobiles, applications sur poste de travail, applications Web), le développement de son prédécesseur Swing étant abandonné (sauf pour les corrections de bogues).

JavaFX contient des outils très divers, notamment pour les médias audio et vidéo, le graphisme 2D et 3D, la programmation Web, la programmation multi-fils etc.

JavaFX propose un grand nombre de types de diagrammes différents : en courbes, en barres, en bulles, en fromage... ils peuvent être très utiles pour afficher des informations calculées à partir d'un grand nombre de données[18]

e-Jfoenix est une API javafx qui dispose de composant enrichis et visuellement plus jolie.

f-Opencsv est une bibliothèque open-source qui gère et manipule les fichiers csv, elle permet leur lecture modification et création

III.3) Interface de l’application

L’interface principale propose quatre fenêtres sur les algorithmes suivants

- L’algorithme K-means qui s’applique sur un nombre de points générés aléatoirement (Figure III.1)
- L’algorithme K-means, (Figure III.2)
- L’algorithme dek means par le vote, (Figure III.3)
- L’algorithme k-means par la théorie des jeux (THJ) (Figure III.4)

Figure III)- 1 interface de kmeans aléatoire

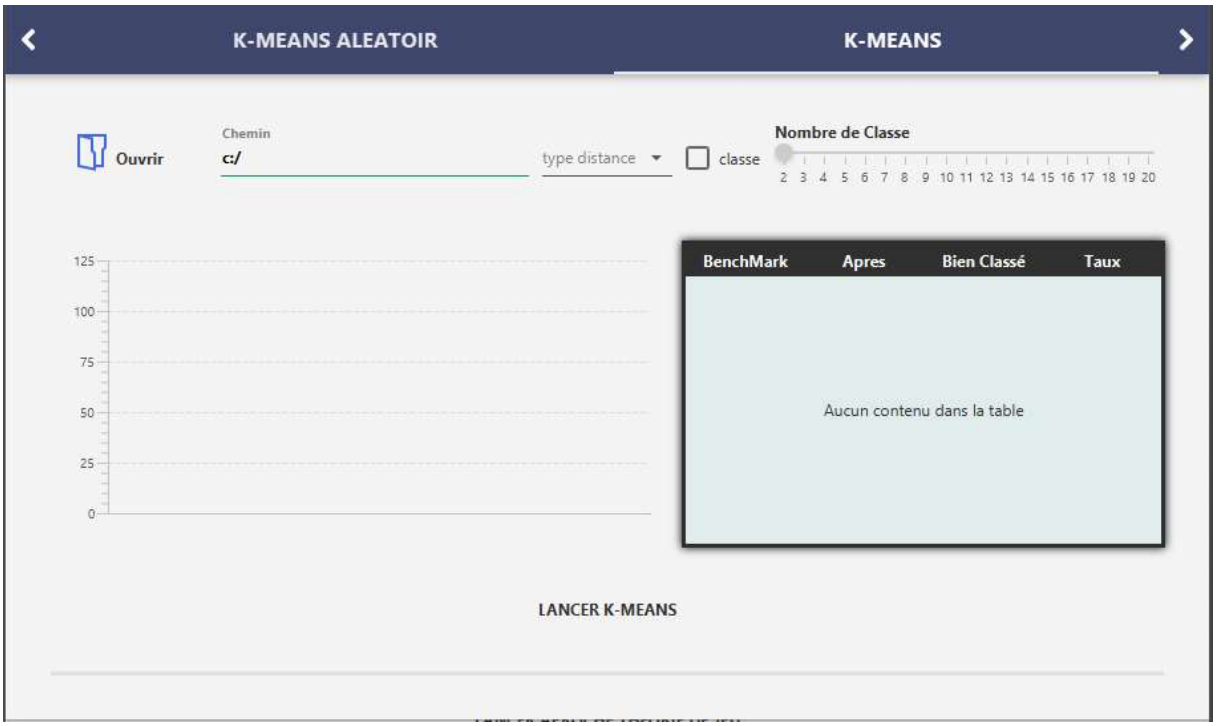

Figure III) 2- interface k-means

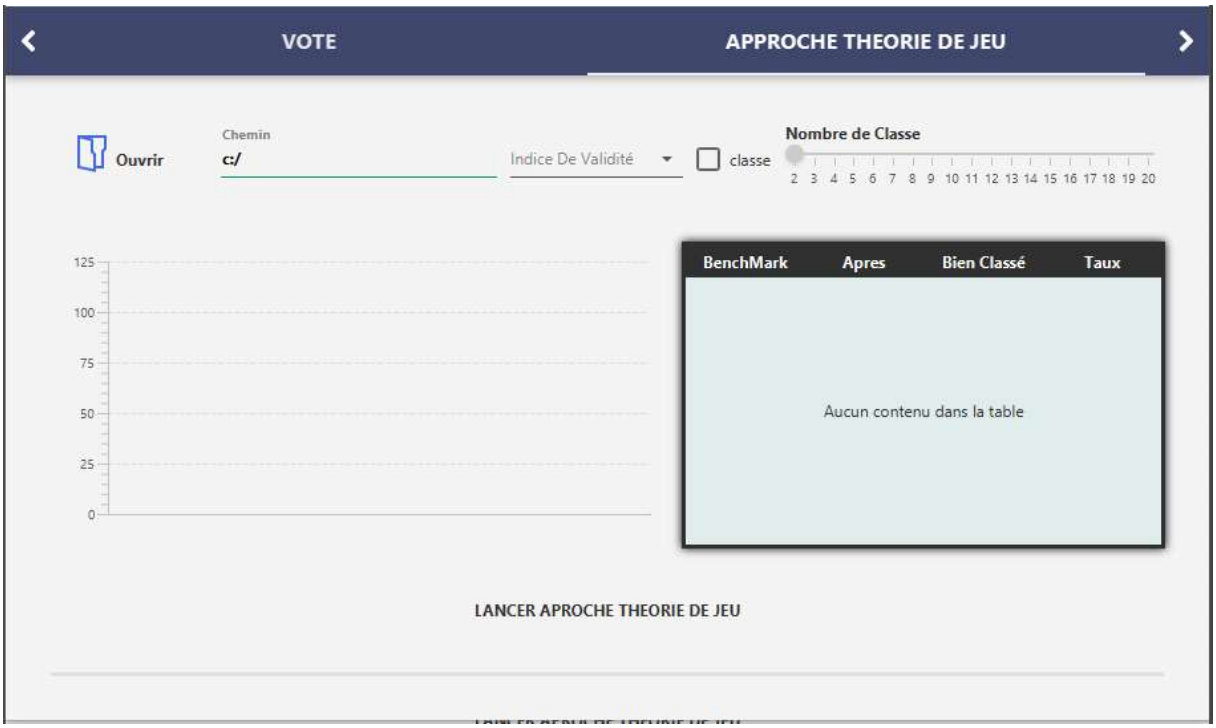

Figure III) 3- interface approche théorie de jeux

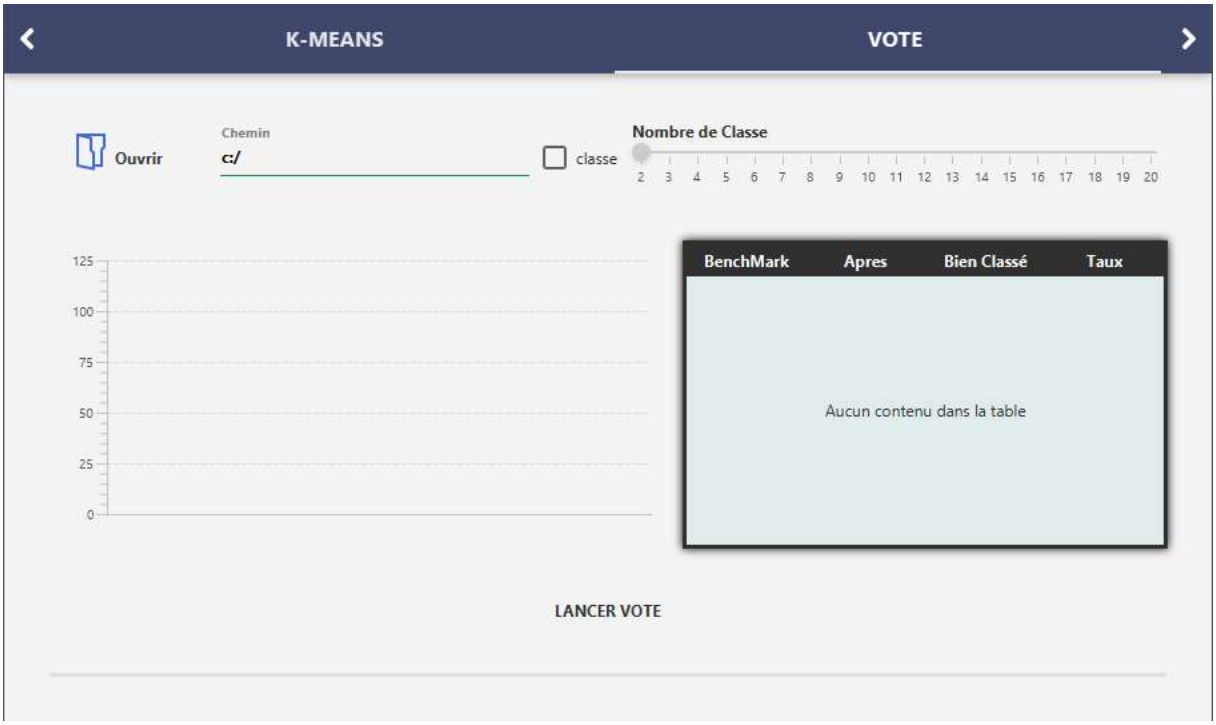

Figure III) 4- interface méthode de vote.

III.4) Les données :

III.4.1) Les benchmark :

Les benchmarks sont des banques de données référentielles, où chaque benchmark a un ensemble de données de nombre et d'attributs différents (voir tableau ci-dessous). Ces benchmarks sont supervisés. Nous pouvons évaluer les résultats de notre algorithme sur ces benchmarks en calculant le taux d'erreurs[28].

	HEART	DIABETES	IRIS	Australian	Satimages	GLASS
Nbre d'objets	270	768	150	690	2000	214
Nbre de classes	2	2	3	2	6	06
Nbre d'attribues	13	8	4	14	36	09
Nbre d'objet dans la classe1	150	500	50	383	397	70
Nbre d'objet dans la classe2	120	268	50	307	211	76
Nbre d'objet dans la classe3	/	/	50	/	237	17
Nbre d'objet dans la classe4	/	/	/	/	470	13
Nbre d'objet dans la classe5	/	/	/	/	224	09
Nbre d'objet dans la classe6	/	/	/	/	461	29

Tableau III) 1-.information sur les benchmarks

III.5) Les résultats :

Les figures suivantes présentent l'exécution des différentes approches, avec les différents paramètres sur la benchmark Diabetes

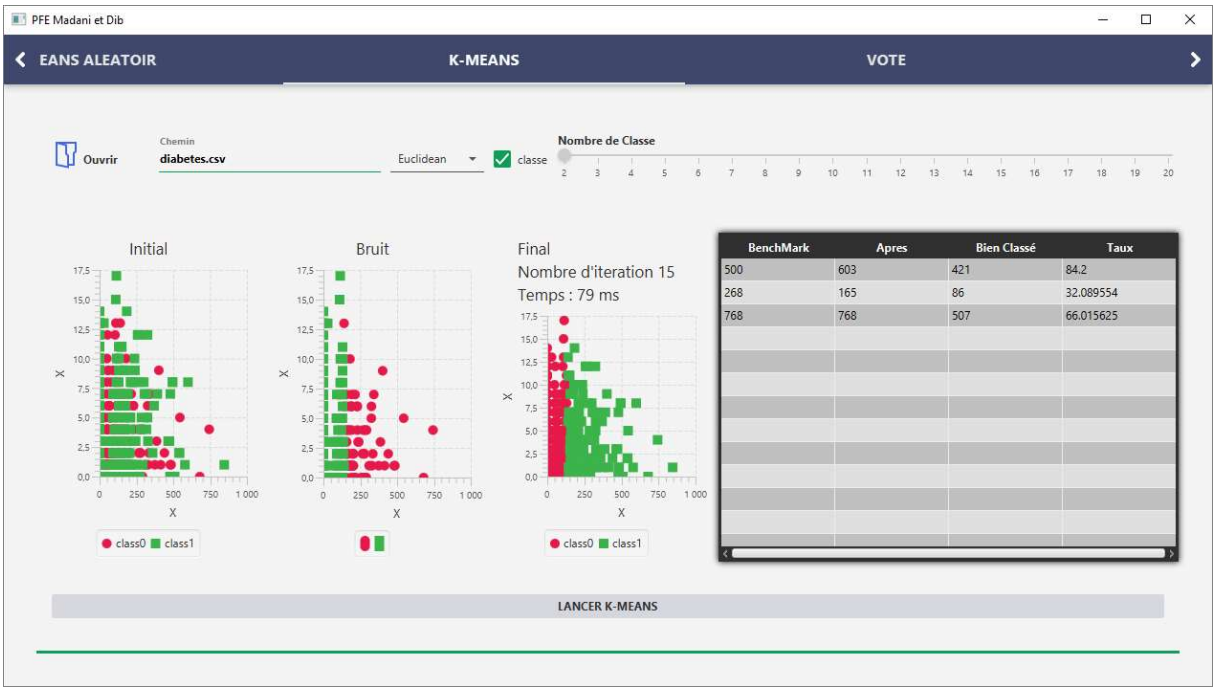

Figure III) 5- résultat du clustering kmeans avec la distance euclidien et k fixé a 2.

Figure III) 6- résultat du clustering kmeans avec distance euclidien, l'indice WB et DB. Cas K variantes

Figure III) 7- résultat du clustering THJ avec l'indice WB et k fixés a 2

Figure III) 8- résultat du clustering THJ avec DB et k fixés a 2

Figure III) 9- Résultat du clustering THJ avec l'indice WB et DB et K variantes

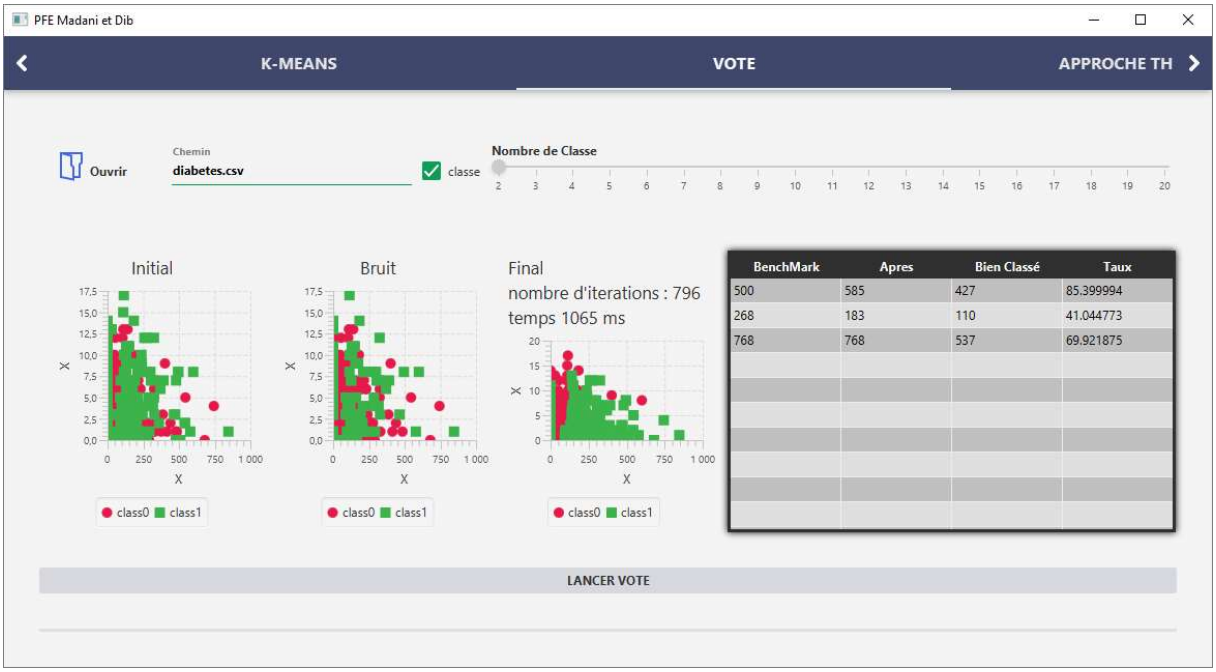

Figure III) 10- Résultat du clustering Vote avec k fixés a 2

Figure III) 11- Résultat du clustering vote avec l'indice DB et WB cas k variante

Afin de vérifier l'efficacité des indices de validité nous avons appliqué les différentes approches de clustering sur les Benchmarks citées ci-dessus, nous avons varié le K de 2 à 10 le tableau III. décrit les caractéristiques des données utilisées et le nombre de cluster renvoyées par les indices de validité.

Benchemarks	K-Means		VOTE		THJ	
	WB	DB	WB	DB	WB	DB
HEART, K=2	5	2	5	2	6	3
IRIS, K=3	6	3	3	3	7	3
GLASS, K= 6	6	3	3	3	7	3
DIABETES, K=2	7	2	2	2	7	2
AUSTRALIANT, K=2	4	2	6	2	5	3
SATIMAGES, K = 6	5	2	3	3	5	3

III. 2 Tableau des caractéristiques des données et le nbr de cluster renvoyé par les indices de validité

On remarque que les deux indices WB et DB renvoient le nombre exact de clusters pour certain Benchmark ou proche pour Benchmark Satimages ou le K = 5 par l'indice WB

Les résultats obtenus montrent l'efficacité de l'indice WB et DB pour extraire le nombre optimal de cluster.

Le tableau III.2 montre les différents taux de classification pour les différents clustering en fixant le K de chaque benchemark

Benchemarks	NUM Classe	K-Means	Vote	Théorie Jeux	
				WB	DB
HEART	1	30.66	73.33	12.66	36
	2	53.33	47.5	14.16	52.49
	Taux	40.74	**61.85**	13.33	43.33
IRIS	1	0	4	30	30
	2	94	56	92	92
	3	0	0	0	0
	Taux	31.33	20	**40.66**	**40.66**
GLASS	1	28.57	20	67.14	67.14
	2	2.63	3.94	0	0
	3	0	0	0	0
	4	0	15.38	23.07	23.07
	5	0	0	11.11	11.11
	6	0	3.44	6.89	6.89
	Taux	10.28	9.34	**24.76**	**24.76**
DIABETE	1	84.2	85.39	34.2	58.8
	2	32.08	41.04	32.08	47.38
	Taux	66.01	**69.92**	33.46	54.81
AUSTRALIAN	1	0	49.34	2.61	12.79
	2	98.37	51.79	71.98	51.46
	Taux	43.76	**50.43**	33.47	30
SATIMAGES	1	0.21	9.76	9.32	9.97
	2	0.89	1.78	0	0.89
	3	0	47.6	0	0
	4	3.31	4.73	0	0
	5	23.62	24.89	2.10	5.06
	6	0	1.06	45.53	42.97
	Taux	3.3	**15.6**	13.09	13.09

Tableau III.3 des résultats du clustering avec le K fixé

Discussion

Benchemark HEART, DIABETE et AUSTRALIAN : par observation des résultas des tableaux on remarque que le clustering vote a donné des meilleurs résultats en terme de taux de classification par rapport au k means et la THJ

Benchemark IRIS, GLASS et SATIMAGES : on remarque que les résultats de la théorie de jeux sont au dessous de la moyenne mais meilleurs par rapport aux taux de classification du k mans et vote

Algorithme	Kmeans		vote		Thj			
Benchemarks	temps	Itérations	Temps	Itérations	WB		DB	
					Temps	Itérations	Temps	Itérations
HEART	43	15	327	348	6090	499	453	21
IRIS	2	12	21	156	996	499	1072	499
GLASS	41	7	250	37	2097	499	3602	499
DIABETE	12	15	345	769	18450	478	1482	21
AUSTRALIAN	40	7	344	781	19438	401	1742	21
SATIMAGES	1618	29	12963	2630	546013	498	52375	21

Tableau III.4 le temps d'execution en ms et le nombre d'itération du clustering

Le tableau III.4 affiche le temps d'exécution et le nombre d'itération de chacune des approche utilisés, et montre que le clustering par théorie des jeux et gourmand en terme de temps surtout pour l'indice de validité WB

III.6) Conclusion :

Nous avons présenté dans ce chapitre notre application et les différents résultats obtenus des différents clustering (le kmeans, la THJ et la méthode de vote).

Les résultats obtenus montrent l'efficacité de l'indice WB et DB pour extraire le nombre optimal de cluster.

Et les résultats obtenus par les deux approches THJ et vote sont prometeurs et ouvrent d'autres perspectives pour améliorer encore les taux de classification.

Conclusion générale et perspectives

Dans ce mémoire, nous avons élaboré une technique de classification non supervisé le "clustering" par partitionnement à base de la théorie des jeux.

Pour cela, nous avons en premier lieu exposé dans le premier chapitre un rappel sur la classification, quelques techniques de classification, répandues dans la littérature et les différents indices de validation du clustering. Parmi toutes les méthodes exposées, nous nous sommes plus particulièrement intéressés au clustering, Un intérêt particulier a été accordé pour l'algorithme Kmeans sur le quel on a introduit le principe de vote.

Le deuxième chapitre de ce mémoire a servi à présenter les notions de base de la théorie des jeux.

Enfin, dans le dernier chapitre, nous avons expliqué notre contribution dans ce domaine qui consiste à l'algorithme kmeans à laquelle nous avons introduit une nouvelle approche de recherche qui se base sur la théorie des jeux.

Les résultats de notre contribution sur les Benchmark pour évaluer leur fiabilité, donnent des résultats assez satisfaisants comparativement aux résultats du k means classique, et révélent aussi l'efficacité des indices de validité pour extraire le nombre optimal de cluster.

On conctate que les résultats de l'approche théorie des jeux peuvent perdre du terrain au profit des autres méthodes, peut-être, moins performantes au niveau des résultats mais avantageuses en termes de complexité calculatoire et en temps de calcul.

Le travail présenté dans ce mémoire, peut avoir un impact sur la suite des travaux de recherches à entreprendre dans l'avenir. On peut citer notamment :

- En étendant notre approche aux autres types d'attributs autres que numériques
- En étendant notre approche sur d'autres types d'indices de validité
- Combiner plusieurs algorithmes de clustering afin d'améliorer les résultats

Les références

Références Bibliographiques :

[4] M. AISSOU BAHAEDDINE, « Analyse des images satellitaires à haute résolution », 2013/2014

[6] P. RAI & S. SINGH, "A survey of clustering techniques, International Journal of Computer Applications" (0975 - 8887) Volume 7- No.12, October 2010.

[7] B. DEVEZE & M. FOUQUIN, « Datamining c4.5 -dbscan, Cours, Ecole d'ingénieurs en informatique »EPITA, France, 2004.

[8] HAMIDOUCHE SADDEK & IDJERAOUI TAYEB, « Clustering : Approche par la théorie des jeux » mémoire de master, Université Mira Abderrahmane de Béjaa 2013

[9] PIERRE-LOUIS GONZALEZ' « méthode de classification », 2008.

[13] COC0QUEREZ (J.) et PHILIPP (S.). « Analyse d'images : filtrage et segmentation ». Masson, 1995.

[14] P. RAI & S. SINGH, "A survey of clustering techniques, International Journal of Computer Applications", (0975 - 8887) Volume 7- No.12, October 2010.

[19] JOHN WILEY & SONS, INC. "Combining Pattern Classifiers: Methods and Algorithms" Second Edition. Ludmila I. Kuncheva.© Published 2014

[20] CHARAFEDDINE KHAMOUDJ, « Proposition et implémentation d'une technique de clustering à base de la théorie des jeux » , mémoire de Magister, 2015

[22] Pattern Clustering using Cooperative Game Theory, 2 Jan 2012, published in Proceedings of Centenary Conference - Department of Electrical Engineering, Indian Institute of Science : 653-658

[23] An incentive compatible multi-armed-bandit crowdsourcing mechanism with quality assurance, Shweta Jain, Sujit Gujar, Satyanath Bhat, Onno Zoeter, Y Narahari, 2014/6/27

[24] Anoop = On the utility of canonical correlation analysis for domain adaptation in multi-view headpose estimation ,KR Anoop, Ramanathan Subramanian, Vassilios Vonikakis, KR Ramakrishnan, Stefan Winkler,2015/9/27

[25] Online topic-based social influence analysis for the wimbledon championships, Varun R Embar, Indrajit Bhattacharya, Vinayaka Pandit, Roman Vaculin, 2015/8/10.

[26] J. HAN & M. KAMBER, Datamining : Concepts and techniques, Management Systems (The Morgan Kaufmann Series in Data Management Systems),MORGAN KAUFFMAN, ISBN 1-55860-901-6, 2006.

[27] S. SABRI, « Application de la théorie des jeux pour la définition, développement et implémentation d'un algorithme de clustering », Mémoire de Magistére, Université de Béjaia, Algérie, 2011.

[29] , FRANCK PICARD, « Classification non supervisée cours , Laboratoire de Biom´etrie et Biologie Evolutive ».

Les références webiographique :

[1] « gestion-des-courriers-électroniques-indésirables »,

-http://communaute-universitaire.univ-rouen.fr/gestion-des-courriers-electroniques-indesirables-441094.kjsp ,Dernière visite : 23/06/2018.

[2] « reconnaissance-optique-caracteres »,https://www.futura-sciences.com/tech/definitions/informatique-reconnaissance-optique-caracteres-11961/, Dernière visite : 07/06/2018.

[3] « une-imagerie-medicale-pour-un-diagnostic-live-des maladies »,https://www.lesechos.fr/thema/030347945424-une-imagerie-medicale-pour-un-diagnostic-live-des-maladies-2089459.php. Dernière visite : 14/05/2018.

[5] « la classification» ,http://dspace.univtlemcen.dz/bitstream/112/1045/4/Memoire.pdf ,

Dernière visite : 04/04/2018.

Les références

[10] « la classification non supervisé », https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-explo-classif.pdf, Dernière visite : 15/02/2018.

[11] « la classification ascendante hiérarchique CAH » https://www.xlstat.com/fr/solutions/fonctionnalites/classification-ascendante-hierarchique-cah, Dernière visite : 15/05/2018.

[12] « Classification par la méthode des nuées dynamiques (k-means) », https://www.xlstat.com/fr/solutions/fonctionnalites/classification-par-la-methode-des-nuees-dynamiques-k-means , Dernière visite : 21/05/2018.

[15] « Cours NetBeans pour débutant en pdf », https://www.cours-gratuit.com/cours-netbeans/cours-netbeans-pour-debutant-en-pdf, Dernière visite : 23/06/2018

[16] « maven» https://www.jmdoudoux.fr/java/dej/chap-maven.htm, Dernière visite : 23/06/2018

[17] « Gérez vos codes source avec Git » https://openclassrooms.com/courses/gerez-vos-codes-source-avec-git, Dernière visite : 23/06/2018

[18] « Introduction à JavaFX » ,http://www.labri.fr/perso/johnen/pdf/IUT-Bordeaux/UMLCours/IntroductionJavaFX-V1.pdf, Dernière visite : 23/06/2018

[21] « les jeux coopératif » ,http://www.occe.coop/~ad82/IMG/pdf/Topo_pour_les_collegues-2.pdf, Dernière visite : 15/04/2018.

[28] « les benchmark » , https://archive.ics.uci.edu/ml/datasets.html, , Dernière visite : 15/04/2018.

Résumé

Résumé

Dans ce travail, notre motivation a été d'étudier la problématique de clustering, qui a un champ d'applications très larges. Donc nous avons proposé une technique de clustering par partitionnement à savoir le k means à base de la théorie des jeux. La théorie des jeux peut être utilisée pour modéliser un système sous forme de jeu, règles, et joueurs. Dans notre approche, nous avons introduit l'indice de validité comme critère d'évaluation d'homogénéité. Les résultats obtenus sont satisfaisants comparativement aux résultats du k means classique

Mots clé : clustering, k-means, théorie des jeux, indice de validité.

Summary

In this work, our motivation was to study the problem of clustering, which has a wide range of applications. So we proposed a partitioning clustering technique namely the K-MEANS based on game theory. Game theory can be used to model a system in the form of games, rules, and players. In our approach, we introduced the validity index as a homogeneity evaluation criterion. The results obtained are satisfactory compared to the results of the classical k means

Key words: clustering, k-means, game theory, validity index.

ملخص: في هذا العمل، كان دافعنا هو دراسة مشكلة التجميع، التي لديها مجموعة واسعة من التطبيقات. لذلك اقترحنا تقنية تقسيم التجميع وهي kmeans بناءً على نظرية الألعاب نظرية الألعاب يمكن استخدامها لنمذجة نظام في شكل الألعاب والقواعد واللاعبين. في نهجنا ، قدمنا مؤشر الصحة كمعيار لتقييم التجانس. النتائج التي تم الحصول عليها مرضية مقارنة بنتائج kmeans التقليدي

الكلمات المفتاحية : التجميع ، k-means، نظرية الألعاب ، مؤشر الصلاحية.