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: *Modèle d'Intelligence et Décision (M.I.D)*

Thème

Comparaison entre les indices de validité pour obtenir le cluster optimal dans le clustering

Réalisé par :

- FEDDANE Esma Nour El Houda
- BRAHIMI Fatima Zohra

Présenté le 04 juillet 2023 devant le jury composé de:

- M^me^ BENMAHDI Meriem Bouchra (Présidente)
- M^me^ CHAOUCHE RAMDANE Lamia (Encadrante)
- M^me^ KAZI TANI Adila (Examinatrice)

Année universitaire : 2022-202

Remerciement

Nous tenons tout d'abord à remercier Dieu le tout puissant et miséricordieux, qui nous a donné la force et la patient d'accomplir ce modeste travail.

Nous tenons à remercier en premier lieu notre encadrante Madame CHAOUCHE RAMDANE pour l'orientation, la confiance et se bonne explication qui ont constitué un apport constitué sans lequel ce travail n'aurait pas pu être mené au bon port.

Nos remerciements s'adressent également aux membres du jury que notre à faits le grand honneur d'évaluer ce travail.

Nous tenons à exprimer nos remerciements à tous nos enseignants qui nous ont enseigné et qui par leurs compétences nous ont soutenu dans la poursuite de nos études.

Enfin, nous également à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

Dédicaces

Nous dédions ce modeste travail à :

Nos chers parents, pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de nos études.

Nos chers frères et chères sœurs, pour leur encouragement, leur appui, leur soutien moral.

Tous nos amis qui nous ont toujours encouragés, et à qui nous souhaite plus de succès.

Table des matières

Introduction générale

La classification des données est une tâche essentielle dans le domaine de l'apprentissage automatique et de l'exploration de données. Elle consiste à attribuer des étiquettes ou des catégories prédéfinies à des instances de données en fonction de leurs caractéristiques ou de leurs propriétés observées. L'objectif est de construire un modèle ou un algorithme capable de généraliser à de nouvelles données non étiquetées, en se basant sur les connaissances acquises à partir des données d'apprentissage.

La classification des données présente de nombreux avantages et applications dans divers domaines. Elle permet de résoudre des problèmes tels que la détection de spam dans les e-mails, la prédiction de maladies à partir de symptômes, la reconnaissance de caractères manuscrits, la détection de fraudes financières, et bien d'autres encore. En classifiant les données, on peut obtenir des informations précieuses et prendre des décisions éclairées en se basant sur des motifs et des relations identifiés dans les données.

Les méthodes de classification utilisent généralement des modèles d'apprentissage automatique tels que les arbres de décision, les réseaux de neurones, les machines à vecteurs de support (SVM) et les algorithmes basés sur les voisins les plus proches (k-plus proches voisins). Ces modèles sont entraînés à partir de données d'apprentissage étiquetées, où les caractéristiques des données sont utilisées pour apprendre des règles ou des motifs qui permettent de discriminer et de classer les instances de données. On appelle cette approche la classification supervisée.

La classification non supervisée, aussi appelée clustering est une technique qui tend à générer à partir d'un ensemble de données non labélisée, des groupes ou des clusters homogènes. Généralement, un bon clustering est synonyme d'une faible inertie intraclusters et une grande inertie inter-clusters. De par sa facilité de mise en oeuvre, le clustering est largement utilisé dans de nombreux domaines, tels que, la fouille de données, la bio-informatique, la reconnaissance des formes et l'indexation des bases d'images. Cependant, la principale limite lors de l'application de ces algorithmes réside

dans le nombre de clusters k fixé préalablement. En effet, à chaque initialisation de ce paramètre peut correspondre une solution différent. Pour surmonter cette limitation, nous proposons dans ce mémoire, l'utilisation de plusieurs indices de validité de clustering conjointement avec l'algorithme de classification non supervisée le plus populaire à savoir le K means. Dans ce travail, nous étudions l'efficacité de certains indices pour déterminer le nombre optimal de clusters.

Le présent mémoire est organisé en trois chapitres :

Le premier chapitre, présente un aperçu général sur les différentes techniques de classification. Les différents critères d'évaluation de la classification seront présentés dans le deuxième chapitre. Le dernier chapitre est consacré aux expérimentations menées sur des données synthétiques en effectuant une comparaison sur plusieurs indices de validité internes avec l'algorithme du K means pour déterminer le nombre optimal de clusters.

Et enfin, nous conclurons ce mémoire par une conclusion générale et quelques perspectives.

Chapitre I **Classification**

Chapitre I Classification

I.1 Introduction

La classification est utilisée pour classer chaque élément de données dans un ensemble de clusters prédéfinis. La classification des tâches d'analyse des données consiste à construire un modèle pour prédire l'étiquette de classification. La classification est une fonction d'exploration de données qui affecte les données d'une collection à des classes cibles. Le but de la classification est de prédire avec précision chaque cas dans les données. Par exemple, des modèles de classification peuvent être utilisés pour identifier les prêts à risque de crédit faible, moyen ou élevé. Les tâches de classification commencent à partir d'ensembles de données avec des affectations de classe connues. Par exemple, un modèle de classification qui prédit le risque de crédit peut être développé sur la base d'observations de prêts de nombreux demandeurs sur une période donnée. La classification de données situées dans un espace de grande dimension est un problème délicat qui se pose dans de nombreuses sciences, où le but général est de pouvoir étiqueter des données en leur attribuant des catégories. Il existe trois types de méthodes : la classification supervisée, la classification non supervisée et la classification semi supervisée. [Kesavaraj, 2013] [Khedairia 2014]

Si l'on dispose d'un ensemble de points étiquetés, on parlera de classification supervisée. Dans le cas contraire, on effectue une classification non supervisée (Clustering en anglais). Par contre la classification semi supervisée utilise un ensemble de données étiquetées et non étiquetées. Dans cette partie, nous exposons les différentes méthodes de classification et nous mettons l'accent sur la classification non supervisée. [far, 2021]

Chapitre I

I.2. Classification supervisée

Lorsque les différentes classes sont connues et que les exemples sont étiquetés avec leur classe respective, on parle de classification supervisée. L'objectif de ce processus est d'utiliser un modèle d'apprentissage pour apprendre à partir d'un ensemble d'exemples (appelé ensemble d'apprentissage) afin de déterminer les règles permettant de prédire la classe des nouveaux exemples. En d'autres termes, il s'agit de découvrir la structure des classes afin de pouvoir l'appliquer à un ensemble de données plus vaste.

Les méthodes supervisées peuvent être classées en deux catégories distinctes :

Les méthodes probabilistes, également appelées méthodes paramétriques, font souvent l'hypothèse d'une distribution gaussienne pour les paramètres statistiques tels que la moyenne, l'écart type, la variance et la covariance. Afin de comprendre la distribution des données, différentes techniques statistiques sont utilisées pour discriminer entre les classes. Parmi les méthodes probabilistes les plus utilisées on cite la distance de Mahalanobis, la méthode Parallelepipedique, la méthode du Minimum de distance et la méthode du Maximum de vraisemblance.

Les méthodes géométriques, également connues sous le nom de méthodes non paramétriques, sont souvent préférées en raison de leur faible coût en termes de temps de calcul. Elles ne tiennent pas compte de la distribution probabiliste. Quelques exemples de ces méthodes incluent l'arbre de décision, la méthode de Sebestien, le K-plus proches voisins (K-ppv), la méthode barycentrique et la méthode elliptique.

I.2.1 Maximum de vraisemblance

La méthode de classification supervisée du Maximum de vraisemblance consiste à choisir le modèle de distribution de probabilité le plus approprié pour chaque classe, estimer ses paramètres à partir des données d'apprentissage, puis utiliser ces paramètres pour calculer la probabilité qu'un nouvel échantillon appartienne à chaque classe et attribuer l'échantillon à la classe avec la plus grande probabilité. [Kharki, 2021]

I.2.2 Arbre de décision

L'arbre de décision est un algorithme qui organise les données en les classant sous forme de branches. À partir d'un nœud racine, chaque donnée suit une direction spécifique en fonction de ses caractéristiques. Cela permet ensuite de prédire les variables de réponse associées aux données. Les points de rencontre dans un arbre de décision sont appelés nœuds, et leurs objectifs sont représentés par les feuilles. Les nœuds servent à définir les règles pour séparer les données en différentes catégories, tandis que les feuilles contiennent les informations elles-mêmes. Prenons l'exemple de la figure I.1 et de la figure I.2. [S01]

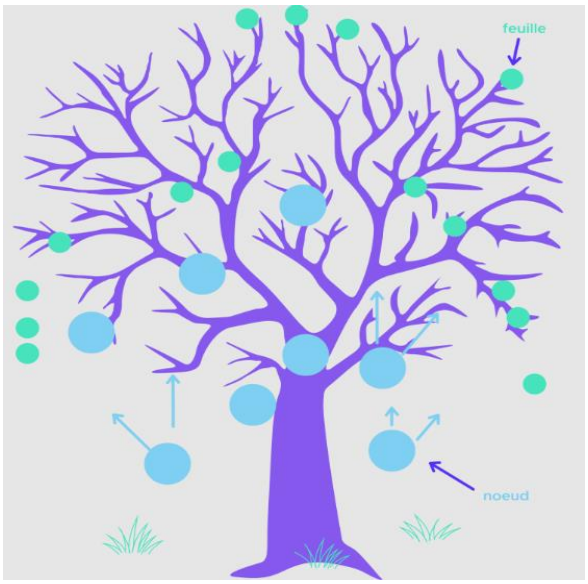

Figure I.1 : Arbre de décision [S02]

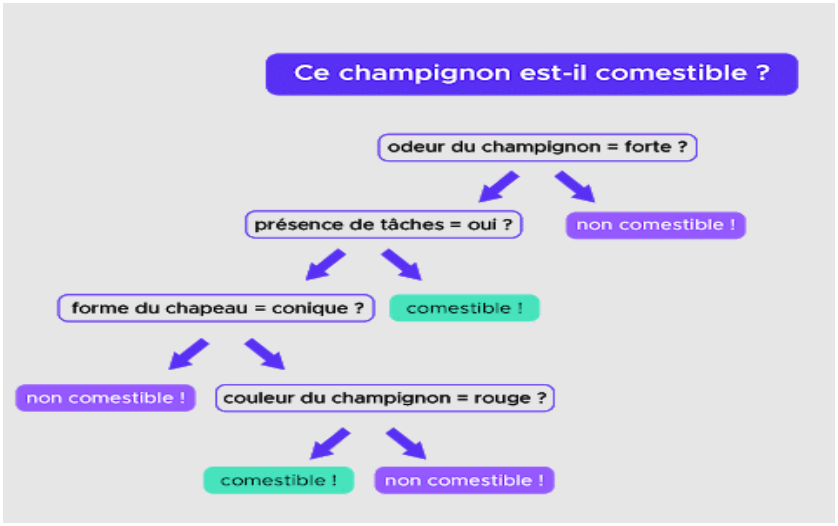

Figure I.2 : Exemple d'arbre de décision [S02]

I.2.3 K Plus Proches Voisins

Le principe général de la méthode k-ppv est de chercher dans un ensemble d'entraînement T. Il contient l'ensemble des individus et leurs classes assignées, les k individus les plus proches de l'individu étant classés. L'individu est alors affecté à la classe majoritaire parmi ces k individus trouvés. Le nombre k est prédéterminé par l'utilisateur. Si k = 1, l'individu est affecté à la classe du plus proche voisin dans l'ensemble T. Une variante de la règle de la majorité consiste à fixer un seuil au-dessus du quel une décision de rejet est prise. Par conséquent, la personne ne peut être affectée à aucune classe. Prenons l'exemple de la figure I.3. Deux dimensions correspondent aux attributs el et e2 et k=3. [Laouamer ,2006]

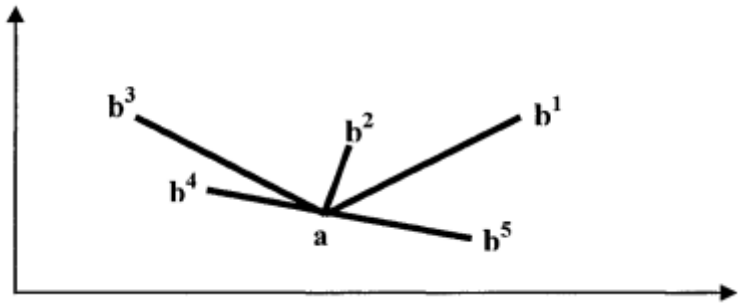

Figure I.3 : Méthode des 3-pvv [Laouamer ,2006]

Dans cet exemple les trois plus proches voisins de a sont b4, b2 et b5, donc a sera affecté à la classe majoritaire parmi ces trois points.

I.3 Classification non supervisée

La classification non supervisée (ou clustering) est une technique d'apprentissage automatique où l'algorithme cherche à découvrir des structures ou des schémas intrinsèques dans les données sans utiliser d'étiquettes de classe préexistantes. Elle regroupe les données en clusters ou en classes similaires sans avoir de connaissances à priori sur les catégories.

I.3.1 Différents type de clustering

Il existe trois façons différentes de formaliser le processus de clustering le clustering dur, le clustering flou et le clustering doux. [Germain, 2010]

Dans le tableau I.1, un exemple des degrés d'appartenance des objets aux clusters est présenté pour les résultats obtenus avec chacune de ces approches (dur, doux et flou).

Le clustering dur (hard-clustering) est la méthode la plus couramment utilisée. Elle consiste à attribuer un objet à une et une seule classe.

Le clustering flou (fuzzy-clustering) spécifie le degré d'appartenance d'un élément à un groupe particulier.

Le clustering doux (soft-clustering) également appelé clustering par recouvrement, propose une attribution rigide de chaque objet à une ou plusieurs classes. Les approches de clustering doux sont assez rares.

	C_1	C_2	C_3
X_1	1	0	0
X_2	0	1	0
X_3	0	0	1
X_4	0	0	1

Exemple de résultat dur

	C_1	C_2	C_3
X_1	1	1	0
X_2	0	1	1
X_3	0	1	1
X_4	0	0	1

Exemple de résultat doux

	C_1	C_2	C_3
X_1	0.9	0.1	0
X_2	0	0.8	0.2
X_3	0	0.7	0.3
X_4	0	0	1

Exemple de résultat flou

Tableau I.1 : Exemple des degrés d'appartenance des objets aux clusters pour un résultat dur, doux et flou [Germain, 2010]

I.3.2 Les méthodes basées sur la distance

La plupart des algorithmes de classification sont basés sur le concept de distance entre les données comme mesure de similarité (ou de dissemblance). Dans ce contexte, les algorithmes de classification tentent souvent d'optimiser une fonction objective qui favorise les clusters compacts et bien séparés. [Sublime, 2016]

On distingue les différentes distances dans le tableau I.2.

distance euclidienne	$d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$
distance de Manhattan	$d(x,y) = \sum_{i=1}^{p}\|x_i - y_i\|$
distance de Minkowski	$d(x,y) = \left(\sum_{i=1}^{p}\|x_i - y_i\|^q\right)^{\frac{1}{q}}$
distance de Canberra	$d(x,y) = \sum_{i=1}^{p}\frac{\|x_i - y_i\|}{\|x_i + y_i\|}$
distance maximum	$d(x,y) = sup_{i\in\{1\ldots p\}}\|x_i - y_i\|$
Hamming distance	$d(x, y) = \sum_i(1 - \delta_{x_i,y_i})P$

Tableau I.2 : Exemples de distances communes [Gueye, 2019] [Sublime, 2016]

I.3.3 Classification hiérarchique

Les algorithmes de clustering hiérarchique adoptent une approche différente en créant une structure de données basée sur un arbre binaire appelé dendrogramme. Une fois que le dendrogramme est construit, il devient possible de choisir automatiquement le nombre approprié de clusters en divisant l'arbre à différents niveaux, offrant ainsi différentes solutions de regroupement pour le même ensemble de données, sans avoir besoin de relancer l'algorithme de clustering. Le clustering hiérarchique peut être réalisé de deux manières distinctes: le clustering ascendant et le clustering descendant. Bien que ces deux approches utilisent le concept de dendrogramme pour organiser les données, elles peuvent produire des ensembles de résultats totalement différents en fonction du critère utilisé lors du processus de clustering. [Reddy ,2014]

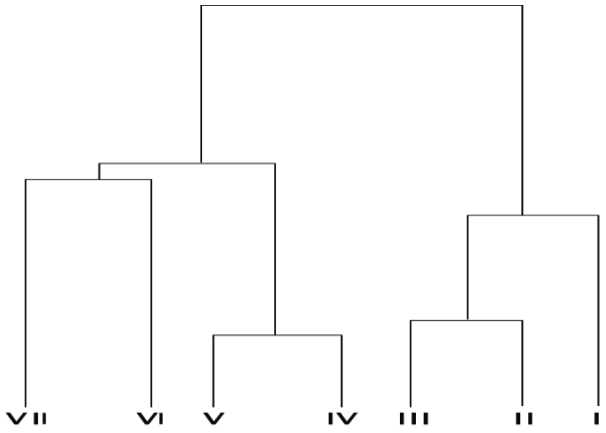

Figure I.4 : Exemple de dendrogramme [S03]

I.3.4 K means

Cet algorithme, également connu sous le nom d'algorithme des centres mobiles, utilise le concept de centres de gravité pour représenter chaque classe (Voir figure I. 5). [MacQueen.1967] Son principe est le suivant :

1. On commence par choisir k centres arbitraires c_1, c_2, ..., c_k, où chaque ci représente le centre d'une classe C_i. Chaque classe C_i est définie comme un ensemble d'individus qui sont plus proches de ci que de tout autre centre.
2. Après cette initialisation, on effectue une deuxième partition en regroupant les individus autour des m_j, qui deviennent alors les nouveaux centres (m_j représente le centre de gravité de la classe c_j, calculé à partir des nouvelles classes obtenues).
3. Ce processus est répété jusqu'à atteindre un état de stabilité où aucune amélioration supplémentaire n'est possible.

Cette méthode est convergente et présente l'avantage d'être efficace en termes de temps de calcul. Cependant, son succès dépend fortement de la partition initiale choisie. Une mauvaise initialisation peut conduire à des résultats sous-optimaux.

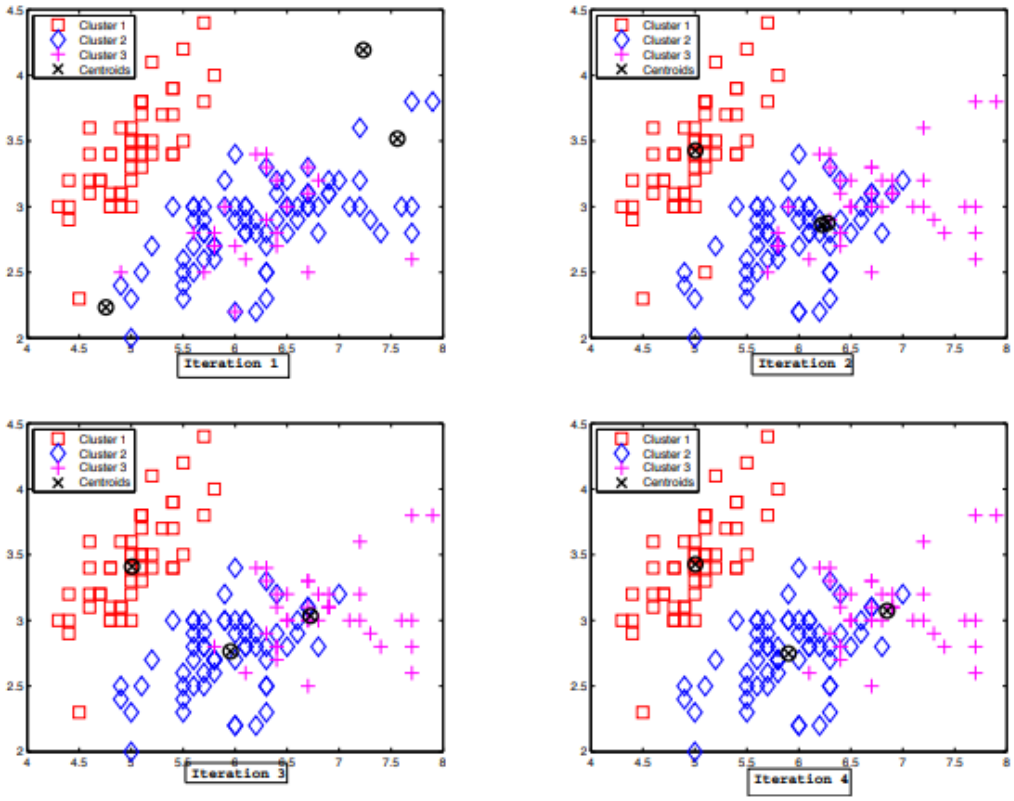

Figure I.5 : Illustration de l'algorithme k-mens [Reddy ,2014]

Cet algorithme vise à minimiser la fonction objective suivante :

$$j = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Ou $\left\| x_i^{(j)} - c_j \right\|^2$ est une mesure de distance choisie entre un point de données $x_i^{(j)}$ et le centre de cluster C_j, qui est un indicateur de la distance des n points de données de leurs données de leurs centres de cluster respectifs.

Les étapes de l'algorithme k-means sont données par :

Entrée : Points de données X, Nombre de clusters k.

Etape 1 : Initialisation de k centroïdes de manière aléatoire.

Etape 2 : Associer chaque point de données sont ainsi divisés en k clusters.

Etape 3 : Recalculez la position des centroïdes.

Répétez les étapes 2 et 3 jusqu'à ce qu'il n'y ait plus de changement dans l'appartenance des points de données.

I.3.5 les méthodes probabilistes

Les méthodes probabilistes de clustering basées sur des modèles ont été largement utilisées et ont montré des résultats prometteurs dans de nombreuses applications, allant de la segmentation d'image, la recognition manuscrite, le clustering de documents, la modélisation thématique à la recherche documentaire. Les approches de clustering basées sur des modèles tentent d'optimiser l'ajustement entre les données observées et certaines modèles mathématique utilisant une approche probabiliste. Ces méthodes sont souvent basées sur l'hypothèse que les données sont générées par un mélange de distributions de probabilité sous-jacentes. En pratique, chaque cluster peut être représenté mathématiquement par une distribution de probabilité paramétrique,

comme une distribution gaussienne. Ainsi, le problème de clustering est transformé en problème d'estimation de paramètre puisque les données entières peuvent être modélisées par un mélange de distributions de composants K Les points de données qui appartiennent le plus probablement à la même distribution peuvent alors facilement être définis comme des clusters. [Reddy ,2014]

- **Mélange Gaussien**

La méthode mélange gaussien utilise l'assignation doux (clustering doux). L'objet mélange gaussien implémente l'algorithme expectation-maximisation (EM) pour l'ajustement des modèles gaussiens. Il peut également dessiner des ellipsoïdes de confiance pour les modèles multi-variés, et calculer le critère d'information bayésien pour évaluer le nombre de grappes dans les données. Une méthode Mélange gaussien ajusté est fournie qui apprend un modèle de mélange gaussien à partir des données du train. Compte tenu des données de test, il peut assigner à chaque échantillon le gaussien il appartient très probablement à l'aide de la méthode de prédiction du mélange gaussien. [S04]

I.3.6 Les algorithmes basés sur densité

Ce type de clustering utilise la densité comme critère au lieu de la distance. On considère un point comme dense s'il a un nombre de voisins supérieur à un seuil donné. De plus, deux points sont considérés comme voisins s'ils se trouvent à une distance inférieure à une valeur prédéterminée. Il s'agit d'un algorithme DBSCAN (Density-Based Spatial Clustering of Applications with Noise) fondé sur la densité. Prenons figureI.7 un exemple de l'algorithme DBSCAN. [S05]

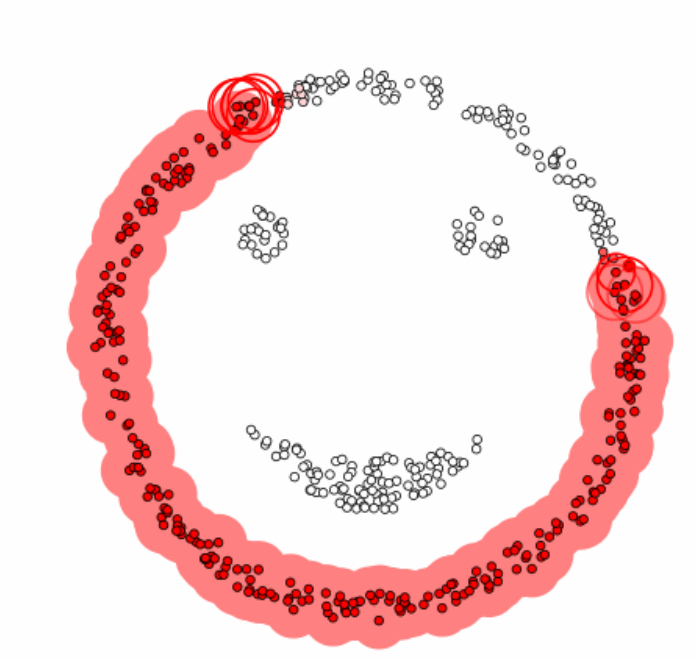

Figure I.6 : Illustration de l'algorithme DBSCAN [S06]

I.3.7 Les méthodes neuronales

Les réseaux de neurones ont été développés pour essayer de créer une représentation mathématique du fonctionnement du cerveau humain. L'idée générale est de créer des unités simples appelées neurones, qui sont capables d'effectuer des calculs élémentaires sur des données numériques. Ces neurones sont ensuite connectés en grand nombre pour former un outil de calcul puissant. Plusieurs modèles ont été proposés dans le but de résoudre des problèmes qui sont difficiles à résoudre avec des méthodes informatiques traditionnelles, voire de rendre le coût de résolution de ces problèmes acceptable selon certains critères. [Norbert, 2021]

I.4 Problèmes et limites du clustering

Malgré l'existence d'un grand nombre de méthodes de clustering et de leur utilisation réussie dans de nombreux domaines, le clustering continue de poser de nombreux problèmes. Ces problèmes sont liés à la fois à l'augmentation de la quantité de données à traiter et à leur hétérogénéité, ainsi qu'au fait que chaque algorithme de clustering nécessite un certain nombre de paramètres, dont le plus important est le nombre de clusters, que l'utilisateur doit définir à l'avance. Un choix inapproprié de ce paramètre peut entraîner des résultats de clustering médiocres.

I.5 Conclusion

Dans ce chapitre, nous avons présenté un aperçu sur les différentes approches utilisées pour la classification : supervisées et non supervisées et les limites du clustering. Pour tenter de résoudre ce problème nous allons voir dans le chapitre suivant les différents critères d’évaluations du clustering.

Chapitre II : les indices de validités

Chapitre II : les indices de validités

II.1 Introduction

L'évaluation de la qualité des résultats de clustering pose une difficulté majeure dans le domaine de la recherche, et cela depuis de nombreuses années, avec l'émergence régulière de nouvelles méthodes. La complexité principale de l'évaluation réside dans la nature intrinsèquement non supervisée du clustering et l'absence de consensus sur ce qui constitue un "bon clustering". Ainsi, l'évaluation d'un résultat de clustering demeure toujours plus ou moins subjective, chaque critère privilégiant un aspect spécifique du bon clustering (forme, compacité, séparation, etc.) par rapport aux autres. Par conséquent, la notion de bon et de meilleur clustering dépend à la fois du critère d'évaluation utilisé et de l'algorithme de clustering lui-même, certains critères favorisant certains algorithmes par rapport à d'autres.

Néanmoins, malgré cette subjectivité relative, il existe un large éventail de critères d'évaluation fréquemment utilisés en apprentissage automatique pour évaluer et comparer les résultats de clustering. La littérature propose plusieurs taxonomies pour ces critères d'évaluation [Halkidi, 2001][Jain,1988][Tan,2005], la plupart d'entre elles définissant trois groupes distincts :

Les indices non supervisés, également appelés indices internes : ils se basent uniquement sur les informations internes des données ainsi que sur les caractéristiques des clusters.

Les indices supervisés, également appelés indices externes : ils évaluent le degré de similitude entre une solution de clustering et une partition connue de l'ensemble de données.

Les indices relatifs, ils constituent une catégorie de critères distincte qui permet de comparer plusieurs résultats de clustering obtenus par un même algorithme. Les indices relatifs utilisent à la fois des critères externes et internes pour sélectionner la meilleure solution parmi plusieurs partitions proposées.

II.2 Indices de validité externe

Les indices externes sont des indices conçus pour mesurer la similarité entre deux partitions. Ils ne considèrent que la distribution des points au sein des différents clusters et ne peuvent pas mesurer la qualité de cette distribution. [Desgraupes, 2013]

Différents critères ont été proposés dans la littérature [Davis, 2006][Fawcett, 2006].Les critères impliquent la comparaison de deux partitions, C1 et C2. Lorsque l'on compare ces partitions, il est important de noter qu'il est généralement impossible d'établir un lien direct entre les clusters des deux partitions, ou dans notre cas, entre les clusters et les classes réelles. En raison de cette limitation, les indices externes ne se basent pas directement sur les classes et les clusters des objets. Au lieu de cela, ils évaluent les associations ou les séparations entre les paires d'objets présentes dans les deux partitions.

aa le nombre de paires d'objets qui sont dans la même classe en C1 et C2,
bb le nombre de paires d'objets qui sont dans des classes différentes en C1 et C2,
ab le nombre de paires d'objets qui sont dans la même classe en C1 mais pas en C2,
et **ba** le nombre de paires d'objets qui sont dans des classes différentes en C1 mais dans la même classe en C2.

Plus **aa** et **bb** sont élevés, plus les deux partitions sont similaires. Nous introduisons maintenant plusieurs critères basés sur ces nombres de paires.

II.2.1 Précision

La précision évalue la probabilité que deux objets soient dans la même classe dans la partition C2 sachant qu'ils sont dans la même classe dans la partition C1. La précision prend ses valeurs entre 0 et 1.

$$\boldsymbol{P = \frac{aa}{aa+ab}}$$

II.2.2 Rappel

Le rappel est la probabilité pour que deux objets soient dans la même classe dans C1 s'ils le sont dans C2. Le rappel prend ses valeurs entre 0 et 1.

$$\boldsymbol{R = \frac{aa}{aa+ba}}$$

II.2.3 Indice de Jaccard

L'indice de Jaccard prend ses valeurs entre 0 et 1 et est égal à 1 si et seulement si les deux partitions C1 et C2 sont identiques.

$$\boldsymbol{Jaccard = \frac{aa+bb}{aa+ab+ba}}$$

II.2.4 Indice de Rand

L'indice de Rand prend des valeurs entre 0 et 1 et est égal à 1 si les deux partitions C1 et C2 sont identiques.

$$\boldsymbol{Rand = \frac{aa+bb}{aa+ab+ba+bb}}$$

II.2.5 F-measure

F- mesure, également connu sous le nom de F1 score, est un autre indice externe basé sur la précision et le rappel. F- mesure prend également ses valeurs dans [0, 1], 1 étant la meilleure valeur.

$$\boldsymbol{F - mesure = \frac{2 \times Precision \times Rappel}{Precision + Rappel}}$$

$$\boldsymbol{F_{\alpha} = \frac{(1+\alpha)P \times R}{\alpha P + R} \quad avec\ \alpha > 0}$$

II.2.6 Indice de kappa

L'indice kappa, est défini par :

$$K = \frac{P_o - P_e}{1 - P_o}$$

Avec

$$P_o = \frac{aa+bb}{aa+ab+ba+bb}$$

$$P_e = \frac{1}{(aa+ab+ba+bb)^2}(aa+ab) \times (aa+ba) + (ab+bb) \times (ba+bb)$$

L'indice Kappa prend ses valeurs entre -1 et 1, 1 signifiant que les partitions C1 et C2 sont identiques.

II.3 Indices de validité interne

Les critères d'évaluation non supervisés reposent sur des informations internes provenant à la fois des données et des clusters. Par exemple, plusieurs d'entre eux se basent sur la distance entre les données et les centroïdes des clusters. Ces indices ont été développés en se fondant sur les principes les plus simples qui définissent un cluster :

(1) Les objets d'un même cluster sont censés être aussi proches que possible les uns des autres.

(2) Les objets appartenant à des clusters différents doivent être bien séparés et aussi éloignés que possible.

Pour évaluer ces critères intuitifs, la plupart des indices adoptent une stratégie consistant à mesurer la distance entre chaque élément de données et un objet représentant les clusters (centroïdes, éléments de données représentatifs, etc.). Ainsi, il est assez simple d'évaluer la compacité et la séparabilité des clusters.

Cependant, en l'absence d'une définition claire de ce qu'est un "bon cluster", chaque indice non supervisé a sa propre méthode de calcul de la compacité et de la séparabilité des clusters, ainsi que d'utilisation de ces deux valeurs pour calculer un critère de qualité final.

Certains de ces critères peuvent être utilisés comme fonctions objectives. L'objectif d'un algorithme de clustering serait alors de trouver une solution qui maximise cette fonction objective. Toutefois, certains critères sont trop coûteux pour être utilisés dans une fonction objective et sont généralement calculés uniquement une fois le processus de clustering terminé. [Pakhira, 2004]

II.3.1 Inerties intra-clusters et inter-clusters

II.3.1.1 Intra-Clusters

La mesure d'inertie intra-cluster évalue le degré de similarité entre les objets qui font partie de la même classe. Elle quantifie les distances qui les séparent par rapport à un point représentant le profil de la classe.

$$\boldsymbol{Intra = \frac{1}{n}\sum_{C\in P}\frac{1}{2n}\sum_{i\in C}\sum_{i\in C} d(i,j)^2}$$

Lorsque les données à l'intérieur des classes sont homogènes, les distances entre ces données et le point représentant la classe sont plus petites. Par conséquent, une valeur basse de l'inertie intra-clusters indique une forte homogénéité des données à l'intérieur des classes.

II.3.1.2 Inter-Clusters

La mesure d'inertie inter-cluster évalue la diversité entre les classes. Elle calcule les distances entre les points qui représentent les profils des différentes classes dans la partition, permettant ainsi de quantifier leur hétérogénéité.

$$\boldsymbol{Inter = \frac{1}{n}\sum_{C\in P} n_C d^2\,(C, C_G)}$$

Avec C représente le centre de la classe et C_G représente le centre du nuage de points.

Lorsque les classes sont plus hétérogènes les unes par rapport aux autres, les distances entre les points représentant les profils des classes sont plus grandes. Par conséquent, une valeur élevée de l'inertie inter-clusters indique une plus grande hétérogénéité entre les classes (voir figure II.1). Il convient de noter que cet indice a tendance à augmenter lorsque le nombre de classes augmente. [Ghribi.2011]

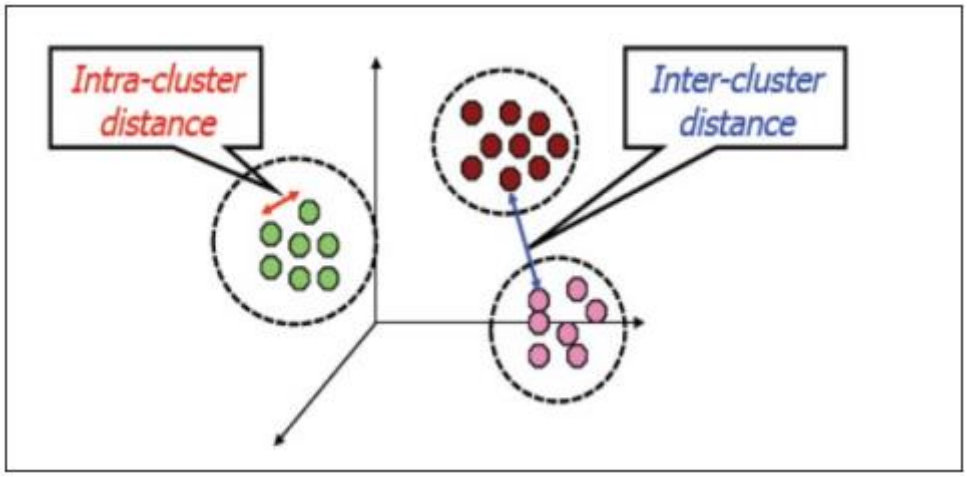

Figure II.1 : Les distances d'intra-cluster et inter-cluster [oubbih, 2016]

II.3.2 Séparabilité et compacité

II.3.2.1 Compacité

La cohésion interne, également appelée "compacité", concerne la formation de partitions connexes avec une structure compacte. L'objectif est de maximiser la similarité entre les objets au sein d'un même cluster. [oubbih, 2016]

- **SSW**

La somme des carrés entre l'objet et le centroïde (en anglais sum-of-squares within clusters) est une mesure de de la compacité [Zhao, 2012], il est défini par la formule suivante :

$$SSW = \sum_{i=1}^{N} \|x_i - c_i\|^2$$

II.3.2.2 Séparation

L'isolement externe, également connu sous le nom de séparation [oubbih, 2016], vise à maximiser la distance entre les points représentant chaque cluster.

- **SSB**

la somme des carrés entre des distances entre centres (en anglais sum-of-squares between) clusters est une mesure de séparation, il est défini par l'équation suivante :

$$SSB = \sum_{i=1}^{M} n_i \|c_i - \bar{X}\|^2$$

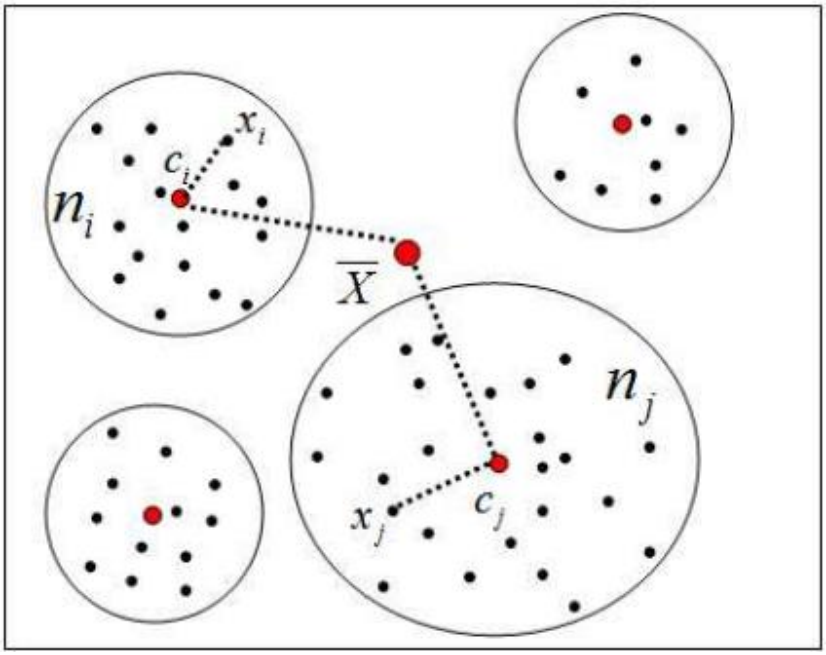

Figure II.2 : Le calcul de SSW et SSB [Zhao, 2012]

II.3.3 Erreurs quadratiques moyennes

L'erreur quadratique moyenne est une méthode simple pour évaluer la qualité des résultats des algorithmes de clustering utilisant des centroïdes. Pour une solution de clustering S avec K clusters, elle peut être calculée selon l'équation où d(.) est une fonction de distance, $|c_i|$ représente le nombre d'éléments dans le cluster c_i, et μ_k est le centroïde du cluster c_k.

$$\boldsymbol{EQM = \frac{1}{\sum_{i=1}^{k}|c_i|}\sum_{k=1}^{K}\sum_{x\in c_k} d(x-\mu_k)^2}$$

Pour considérer un résultat de clustering comme étant de bonne qualité, il est préférable que l'erreur quadratique moyenne soit aussi réduite que possible. [Sublime, 2016]

II.3.4 Indice de Ball-Hall

La dispersion moyenne d'un cluster est calculée en prenant la moyenne des distances au carré entre les points de cluster et leur centre de gravité. L'indice de Ball-Hall est utilisé pour évaluer la dispersion moyenne de tous les clusters. [Desgraupes, 2013]

$$\boldsymbol{C = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i\in I_k}\left\| M_i^{\{k\}} - G^{\{k\}}\right\|^2}$$

Si tous les clusters ont la même taille N/K, la somme mentionnée précédemment peut être simplifiée à :

$$\boldsymbol{BH = \frac{SSW}{N}}$$

II.3.5 Indice Calinski-Harabasz

L'indice Calinski-Harabasz c'est un indice populaire intégrant les mesures SSW, SSB. Il est défini comme :

$$\boldsymbol{CH = \frac{SSB/(K-1)}{SSW/(N-K)} = \frac{N-K}{K-1}\frac{SSB}{SSW}}$$

Avec K le nombre des clusters.

Une valeur forte de l'indice CH indique un clustering de bonne qualité.[Lallich, 2015]

II.3.6 Indice de Dunn

L'indice de Dunn permet d'évaluer à la fois la séparabilité et la compacité des clusters. Il est formulé comme suit :

$$Dunn(C) = \frac{\min_{i=1,..,k}\left(\min_{j=1,...,k,j\neq i}\left(D_s(C_i,C_j)\right)\right)}{\max_{i=1,..,k}\left(D_i(C_i,C_i)\right)}$$

où $D_s(C_i,C_j)$ correspond à la distance entre les clusters C_i et C_j, et définie comme la distance minimale entre les objets des deux clusters.

$D_i(C_i,C_i)$ correspond à la distance maximale entre deux objets du cluster .

Des valeurs faibles indiquent une compacité forte et une forte séparation des clusters. [Germain, 2010]

II.3.7 Indice de Davies Bouldin

L'indice Davies et Bouldin (DB) mesure la compacité et la séparabilité des clusters en se basant sur le calcul de la similarité moyenne entre les clusters. Il est défini comme :

$$DB(C) = \frac{1}{k}\sum_{i=1}^{k} \max_{j=1,..,k,i\neq j}\left(\frac{S(C_i)+S(C_j)}{D(\mu_i,\mu_j)}\right)$$

Où $D(\mu_i, \mu_j)$ est la distance entre les centroïdes de C_i et C_j

$S(C_i)$ est la distance moyenne entre chaque objet de Ci et son centroïde μ_i.

μ_j est la distance entre les clusters de C_i et C_j

$$S(C_i) = \frac{1}{|C_i|}\sum_{X\in C_i} d(x,\mu_i)$$

Plus les centroïdes sont compacts, plus la distance moyenne au centroïde ($S(C_i)$) est petite. Plus les clusters sont éloignés les uns des autres, plus la distance ($d(\mu_i, \mu_j)$) entre les clusters est grande.

Une faible valeur pour l'indice Davies Bouldin indique un bon clustering. [Germain, 2010]

II.3.8 Indice de Hartigan

L'indice de Hartigan est une règle empirique heuristique. Il est également basé sur la mesure SSW, il est défini par l'équation suivante :

$$Hi = (\frac{SSW_K}{SSW_{K+1}} - 1)(N - K - 1)$$

Une valeur maximale de l'indice Hartigan [Sanka, 2021] indique un clustering de bonne qualité. [Hennig, 2015]

II.3.9 Indice de WB

L'indice WB basée sur les mesures SSW et SSB. Il est donné par l'équation:

$$WB = K \times \frac{SSW}{SSB}$$

Avec K le nombre des clusters.

Une valeur faible de l'indice WB indique un clustering de bonne qualité.

II.3.10 Indice de Bayesian information Criterion (BIC)

L'indice BIC [Farhi, 2017] est basé, en partie, sur la fonction de vraisemblance et il est formulé comme suit :

$$BIC = \sum_{i=1}^{M} \left(n_i \, log \, \frac{n_i}{N} - \frac{n_i \times D}{2} log(2\pi) - \frac{n_i}{2} log \, \Sigma_i \right)$$

Et

$$\Sigma_i = \frac{1}{N-M} \sum_{j=1}^{n_i} \left\| x_j - c_i \right\|^2$$

où c_i représente le i iéme cluster,

n_i la taille de celui-ci

x_j le j iéme point dans le cluster c_i. . [Farhi, 2017]

II.3.11 Indice de Silhouette

L'indice Silhouette considérait pour chaque point M_i, sa distance moyenne à chaque cluster. On définit la distance moyenne dans le groupe a(i) comme la distance moyenne du point M_i aux autres points du cluster auquel il appartient:

si $M_i \in C_k$, on a donc

$$a(i) = \frac{1}{n_k - 1} \sum_{i' \in I_k, i' \neq i} d(M_{i'}, M_i)$$

D'autre part, évaluons la distance moyenne σ(M_i, $C_{k'}$) de Mi aux points de chacun des autres clusters $C_{k'}$.

$$\sigma(M_i, C_{k'}) = \frac{1}{n_{k'}} \sum_{i' \in I_{k'}} d(M_i, M_{i'})$$

Dénonçons aussi par b(i) la plus petite de ces distances moyennes :

$$b(i) = min_{k' \neq k}\, \sigma(M_i, C_{k'})$$

La valeur k' qui réalise ce minimum indique le meilleur choix pour réaffecter, si nécessaire, le point M_i à un autre cluster que celui qu'il appartient à

$$S(i) = \frac{b(i) - a(i))}{max(a(i), b(i))}$$

L'indice Silhouette prend des valeurs comprises entre -1 et 1, une valeur proche de 1 indique que le point Mi est affecté au cluster de droite alors qu'une valeur proche de -1 indique que le point devrait être affecté à un autre cluster.

La moyenne des largeurs de silhouette pour un cluster donné C_k est appelé le cluster silhouette moyenne et est indiquée par s_k :

$$s_k = \frac{1}{n_k} \sum_{i \in K_k} s(i)$$

L'indice de silhouette global est la moyenne des silhouettes moyennes à travers tous les clusters. [Desgraupes, 2013]

$$C = \frac{1}{k} \sum_{k=1}^{K} s_k$$

II.3.12 Indice de WSJI

L'indice WSJI permet une combinaison linéaire de compacité floue moyenne et séparation pour mesurer le clustering, qui est défini par l'équation suivante. [Farhi, 2017]

$$WSJI\ (K) = Scat\ (K) + \frac{Set(K)}{Set(k_{max})}$$

Et

$$Scat\ (K) = \frac{\frac{1}{k}\sum_{i=1}^{k}\|\sigma(z_i)\|}{\sigma(x)}$$

$$Set\ (K) = \frac{D_{max}}{D_{min}}\sum_{i=1}^{k}\left(\sum_{i=1}^{k}\|z_i - z_k\|^2\right)^{-1}$$

$$D_{max} = max\ \{\|z_i - z_k\|\}$$

$$D_{min} = min\ \{\|z_i - z_k\|\}$$

II.3.13 Indice Xie-Beni

L'indice Xie-Beni est défini comme un rapport de la variation totale à la séparation minimale des clusters, qui est donné par l'équation suivante. [Farhi, 2017]

$$XB = \frac{1}{N}\frac{\sum_{i=1}^{K}\sum_{j=1}^{N}(U_{ij})^m\|x_i - c_i\|^2}{\min\limits_{l\neq i}\|c_l - c_i\|^2}$$

II.4 Conclusion

Nous avons présenté dans ce chapitre, différents indices internes et externes permettant l'évaluation des résultats de la classification. Parmi ces indices, on s'intéresse aux indices de validité internes afin d'obtenir le nombre optimal de clusters d'une classification non supervisée. Dans le chapitre suivant nous allons donner les différents résultats et discussions de nos expérimentations sur les différents jeux de données.

Chapitre III : Application

Chapitre III : Application

III.1 Introduction

Pour s’assurer de la validité et de la pertinence des indices internes, plusieurs expériences ont été réalisées sur différents jeux de données, tous les résultats des expériences ont été obtenus à l’aide du logiciel SPYDER et sont exécutés sur un PC avec processeur Ryzen7 PRO 5850U avec 16 GO de RAM.

III.2 Environnement de travail

Spyder est un environnement de développement gratuit et open-source conçu spécifiquement pour les ingénieurs et les analystes de données travaillant avec Python. Il offre des fonctionnalités uniques telles que des fonctions avancées d'édition, d'analyse, de débogage et de profilage, ce qui en fait un outil de développement complet. Spyder propose également une exploration interactive des données, une exécution interactive, une inspection approfondie et de superbes capacités de visualisation, faisant de lui un choix idéal pour les professionnels utilisant des packages scientifiques.

Chapitre III

III.3 Data set

Dans cette section, nous proposons d'évaluer et de comparer le comportement des dix indices de validité à savoir les indices Calinski and Harabas, Hartiga, Ball and Hall, Dunn, Silhouette, Davies Bouldin, Xie Beni, WSJI, WB et Bayesian Information Criterion. Pour cela nous avons choisi onze données artificielles (Clustering basic benchmark), les données S1, S2, S3 et S4 qui présentent des taux de recouvrement différents, les données Unbalance, A1, A2 et A3 présentant un nombre différent de clusters et enfin les données Iris, Glass, Dim032 présentant des dimensions différentes.

L'intégralité des données (datasets) peut être obtenue à partir de la page web du SIPU [http://cs.uef.fi/sipu/datasets].

III.3.1 A sets

Ces ensembles 2D contiennent des clusters sphériques avec le nombre de clusters k=20, 35 et 50, de sorte que la taille de cluster (150), l'écart (1402), le chevauchement (20%) et la dimension restent constants. Les ensembles sont des sous-ensembles les uns des autres : A1⊂ A2 ⊂A3 (voir figures III.1, III.2 et III.3).

Figure III.1 : A1 avec N=3000, k=20

Figure III.2 : A2 avec N= 5250, k= 35

Figure III.3 : A3 avec N=7500, k=50

III. 3.2 S sets

L'ensemble de données à deux dimensions qui se compose de 5000 points représentant 15 clusters et la même distribution gaussienne avec un chevauchement croissant entre les clusters. Ces ensembles contiennent des clusters gaussiennes chevauchement (séparation des clusters) de 9 % à 44 %. La plupart clusters sont sphériques. Sauf quelqu'un. Le dernier ensemble (S4) présente un fort chevauchement (voir figures III.4, III.5, III.6 et III.7).

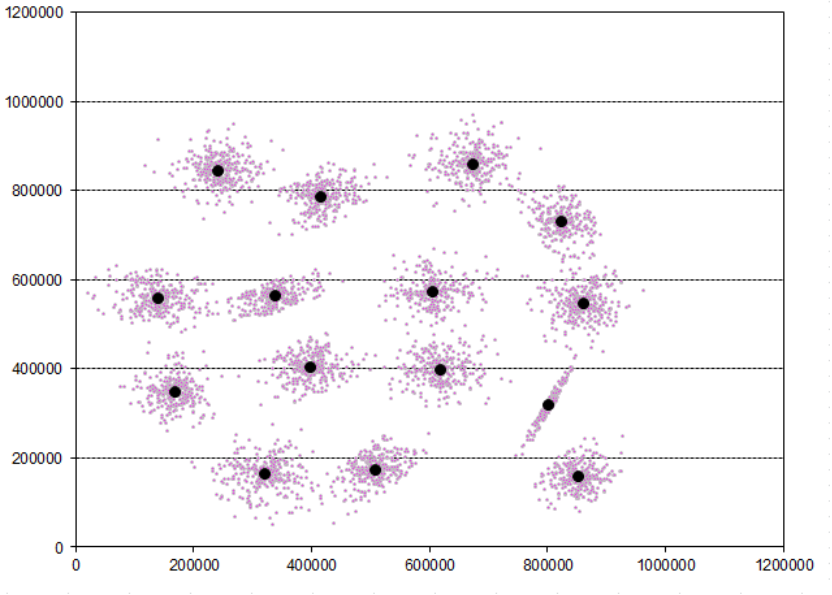

Figure III.4 : S1 avec N=5000, k=15

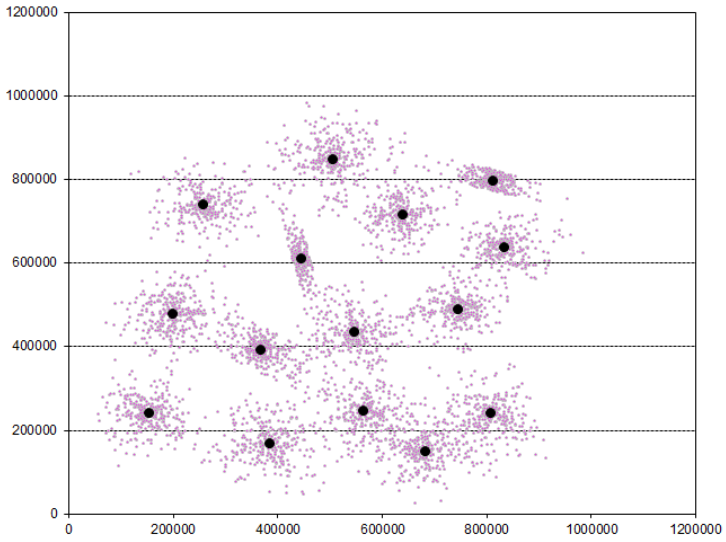

Figure III.5 : S2 avec N=5000, k=15

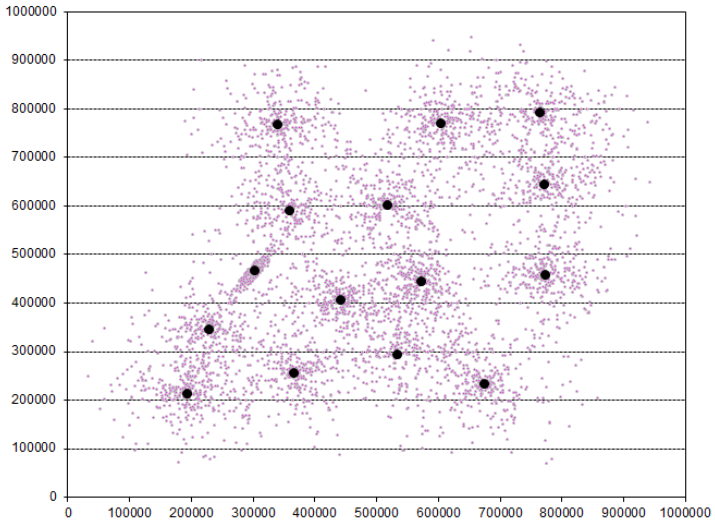

Figure III.6 : S3 avec N=5000, k=15

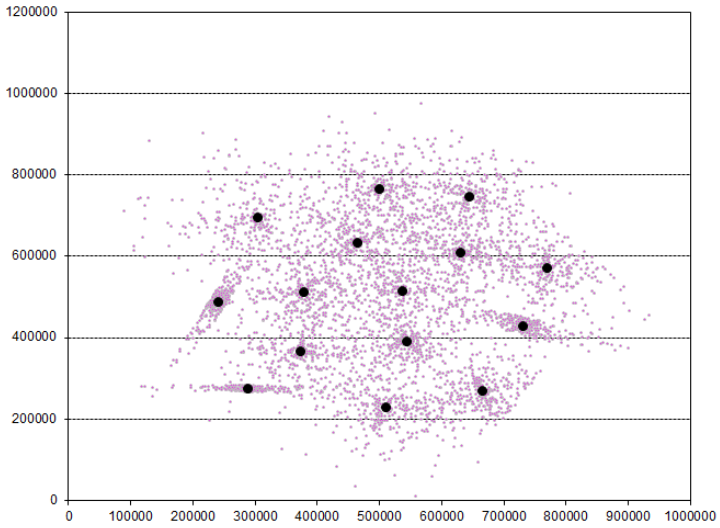

Figure III.7: S4 avec N=5000, k=15

III.3.3 DIM sets

Cet ensemble contient des clusters bien séparés dans un espace de grande dimension avec une dimension de 32. Les points de chaque cluster sont aléatoires et échantillonnés de la distribution gaussienne (voir figure III.8).

Figure III.8 : Dim032 avec N= 1024, k=16

III.3.4 Unbalance sets

L'ensemble de données comprend huit clusters dans deux clusters bien séparés. Les trois premières clusters sont denses avec 2000 points chacune. Les cinq autres clusters sont clairsemées avec 100 points chacune (voir figure III.9).

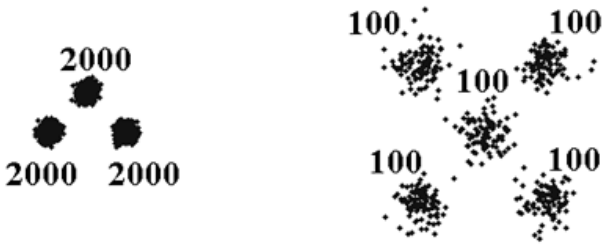

Figure III.9 : Unbalance avec N=6500, k=8

III.3.5 Iris sets

L'ensemble de données [S01] comprend 50 échantillons de trois espèces d'iris (Iris Setosa, Iris virginica et Iris versicolor). Quatre caractéristiques ont été mesurées pour chaque échantillon : la longueur et la largeur des sépales et des pétales, en centimètres (voir figure III.10)

Figure III.10: Iris avec N=150, K=3

III.3.6 Glass sets

L'ensemble de données Glass est plus réaliste avec 214 instances et 9 attributs. Chaque instance représente un morceau de verre, et sa classe est le type de verre. Il existe 6 types possibles, correspondant à différents procédés de fabrication du verre (voir figure III.11).

Figure III.11 : Glass avec N=214, k=6

III.4 Résultats Expérimentaux

Nous proposons une comparaison entre plusieurs indices de validité sur plusieurs dataset en les combinant avec l'algorithme k-means afin de déterminer le nombre optimal de clusters. La méthode consiste à faire varier le nombre de clusters dans un intervalle prédéfini $[k_{min}, k_{max}]$ avec $k_{min} = \boldsymbol{2}$ et k_{max} est donné par la règle empirique = $\sqrt{\frac{N}{2}}$ [Kodinariya et Makwana, 2013].

Le tableau III.1 présente la liste des dix indices utilisés dans notre comparaison.

Nom de l'indice	Cluster optimal
Calinski and Harabas (CH)	Max
Hartiga (H)	Max
Ball & Hall (BH)	Max
Dunn	Max
Silhouette (SIL)	Max
Davies Bouldin (DB)	Min
Xie Beni (XB)	Min
WSJI	Min
WB	Min
Bayesian information Criterion (BIC)	Min

Tableau III.1 : La liste des indices utilisés.

III.5 Résultats et Discussions

Les figures III.12, III.13 et III.14 illustrent les valeurs des indices CH, Sil et DB respectivement sur un intervalle du nombre de clusters k_{min} = 2 et k_{max} = 50 pour les ensembles de données S1, S2, S3 et S4. Nous pouvons constater que la valeur de ces indices indiquent le bon nombre de clusters à savoir 15 clusters pour les données S1- S4 (voir figures III.15, III. 16, III.17 et III.18).

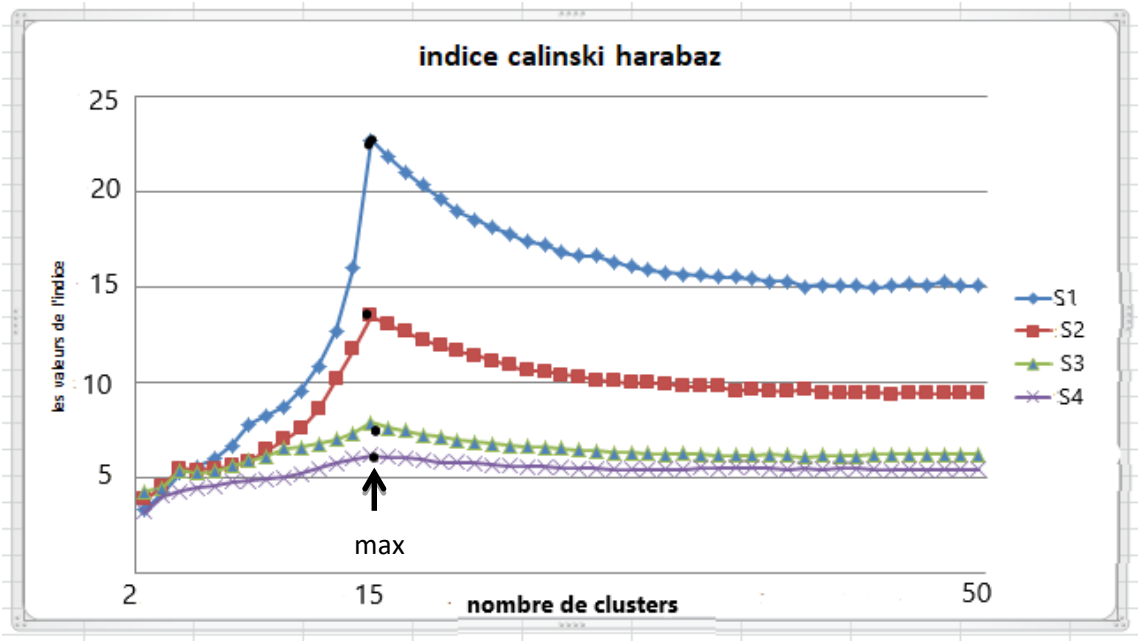

Figure III.12 : Les valeurs de l'indice ***CH*** en fonction du nombre de clusters pour les ensembles de données S1-S4.

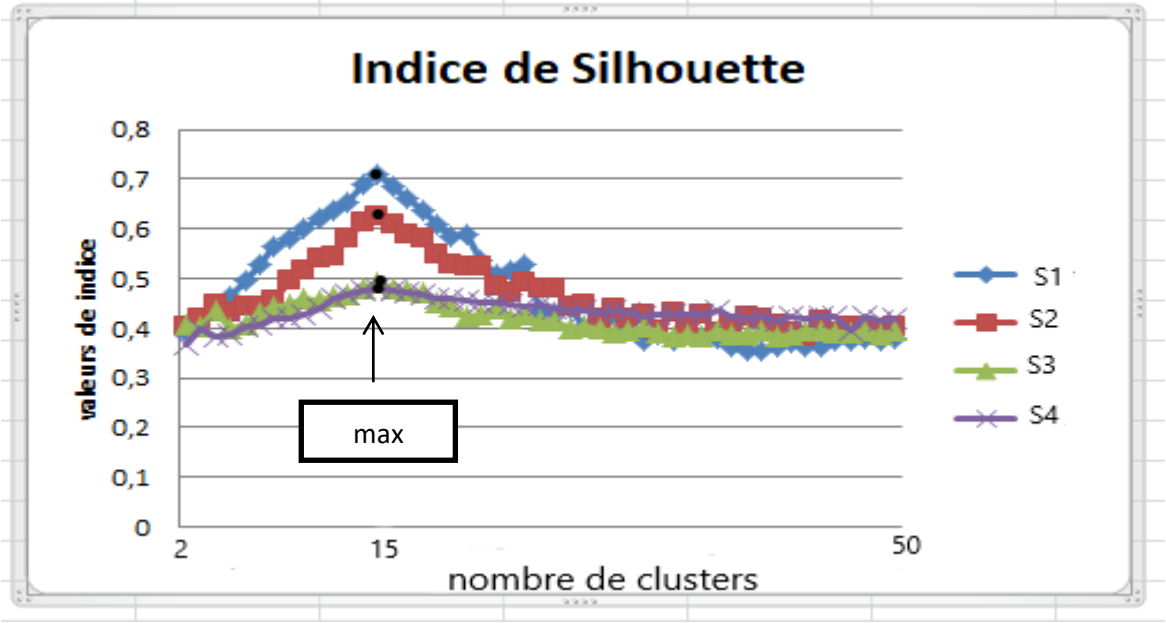

Figure III.13 : Les valeurs de l'indice ***Sil*** en fonction du nombre de clusters pour les ensembles de données S1-S4.

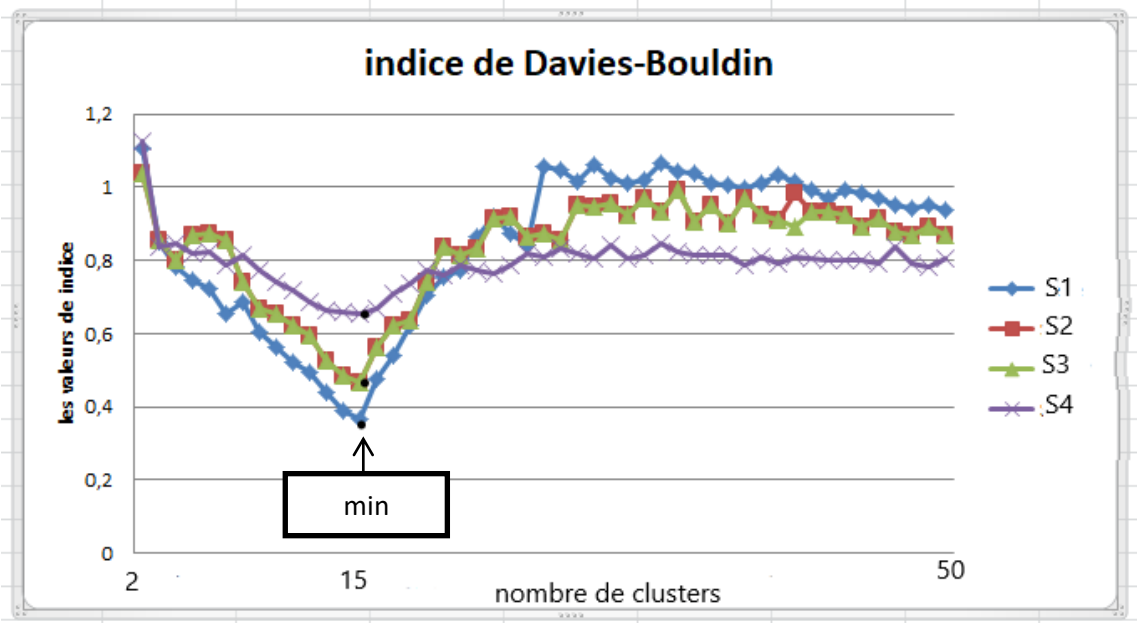

Figure III.14 : Les valeurs de l'indice ***DB*** en fonction du nombre de clusters pour les ensembles de données S1-S4.

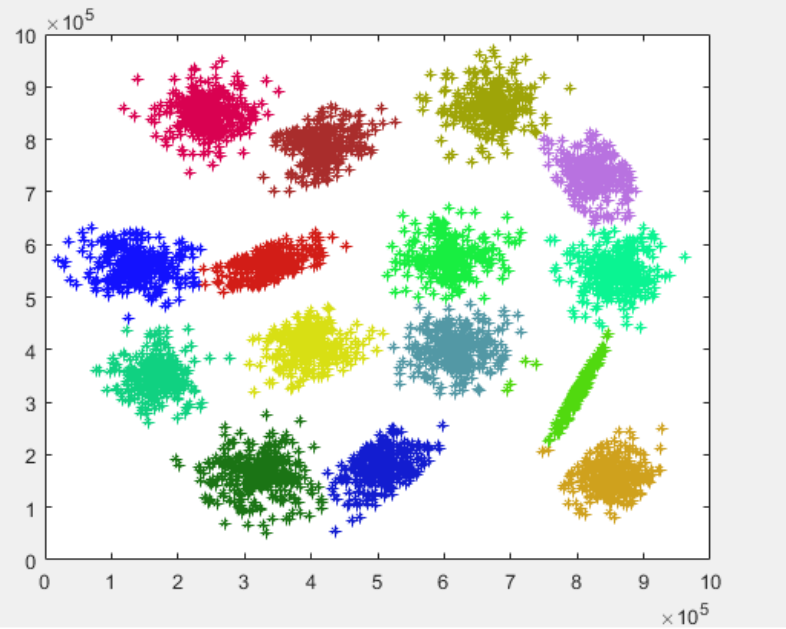

Figure III.15 : Les résultats du K means sur S1 (15 clusters)

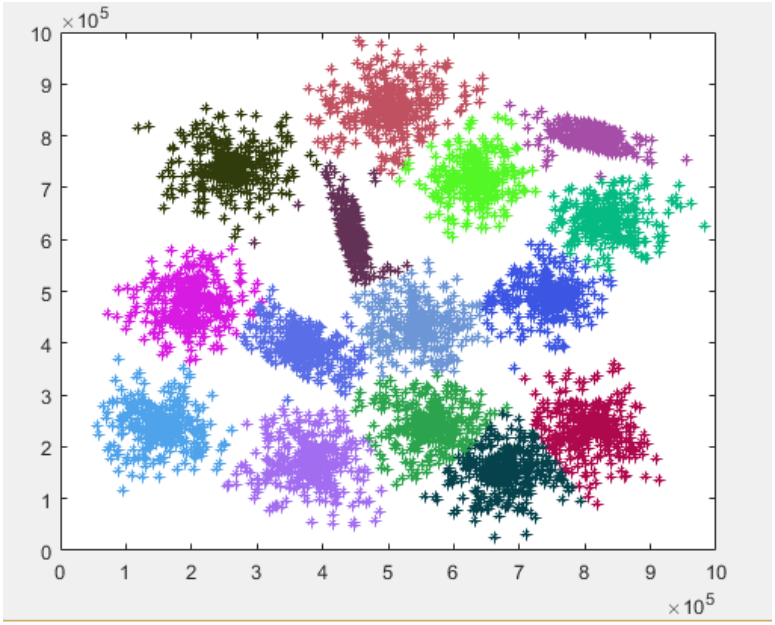

Figure III.16 : Les résultats du K means sur S2 (15 clusters)

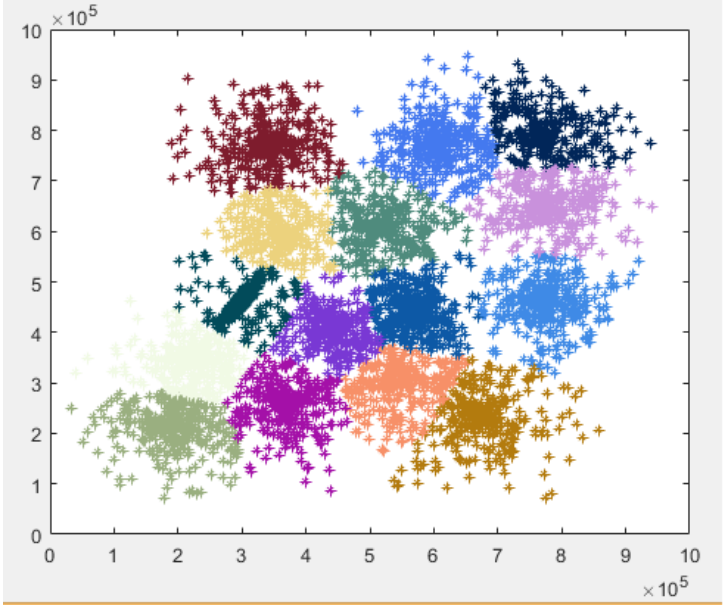

Figure III.17 : Les résultats du K means sur S3 (15 clusters)

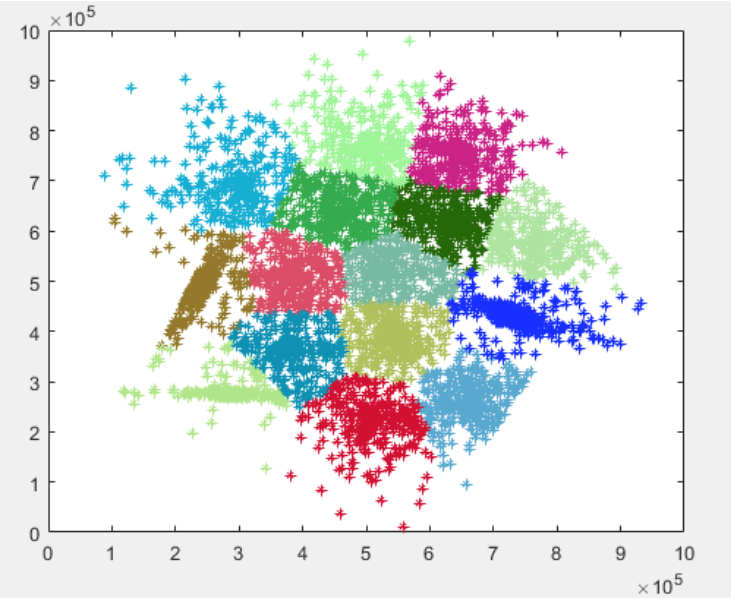

Figure III.18 : Les résultats du K means sur S4 (15 clusters)

D'après le tableau III.2, nous pouvons constater pour les indices SIL, DB, CH, DUNN, BIC, WB les résultats suivants :

- **L'indice Silhouette** retourne le nombre exact de clusters pour les datasets *Dim032, A1, A3 S1 S2, S3, S4 (valeur en vert souligné)* et la valeur proche du nombre de cluster pour *A2 (valeur en bleu)*.
- **L'indice Davies-Bouldin** retourne le nombre exact de clusters pour les datasets *Dim032, A1 S1, S2, S3, S4* et des valeurs proches pour *Iris, Glass, A2.*
- **L'indice Calinski-Harabasz (CH)** retourne le nombre exact de clusters pour les datasets *Iris, Dim032, Unbalance, A1, A2, A3, S1, S2, S3, S4*.
- **L'indice DUNN** retourne le nombre exact de clusters pour les datasets *A1, S1* et des valeurs proches pour *Iris, Dim032.*
- **L'indice BIC** retourne le nombre exact de clusters pour *Unbalance* et la valeur proche pour *Dim032*.
- **L'indice WB** retourne le nombre exact de clusters pour les datasets *Dim032, Unbalance, A1, S1, S2, S3, S4* et des valeurs proches pour *Iris, A2.*

données	NB de cluster	Dimension	Taille	SIL	DB	CH	DUNN	BIC	WB
Iris	3	4	150	6	2	3	4	6	6
Glass	6	9	214	4	7	2	2	10	7
Dim032	16	32	1024	16	16	16	17	17	16
Unbalance	8	2	6500	4	4	8	2	8	8
A1	20	2	3000	20	20	20	20	39	20
A2	35	2	5250	36	34	35	43	50	36
A3	50	2	7500	50	47	50	54	61	53
S1	15	2	5000	15	15	15	15	50	15
S2	15	2	5000	15	15	15	39	50	15
S3	15	2	5000	15	15	15	31	50	15
S4	15	2	5000	15	15	15	37	50	15

Tableau III.2 : Résultats des indices de validité SIL, DB, CH, DUNN, BIC, WB.

D'après le tableau III.3, nous pouvons voir pour les indices BH, XB, WSJI, Hi:

- **L'indice Ball-Hall** retourne la valeur proche pour *Glass*.
- **L'indice Hartigan (HI)** retourne le nombre exact de cluster pour *Unbalance* et des valeurs proches pour *Iris* et *Glass*.
- **L'indice WSJI** retourne le nombre exact de clusters pour les datasets *S1*, *S3* et des valeurs proches pour *Iris*, *Dim032*, *A1*, *A2*, *A3*, *S2*.
- **L'indice Xie-Beni** ne retourne aucune valeur exact ni proche.

données	**NB de cluster**	**Dimension**	**Taille**	**BH**	**XB**	**WSJI**	**Hi**
Iris	**3**	**4**	**150**	6	6	2	2
Glass	**6**	**9**	**214**	5	10	4	5
Dim032	**16**	**32**	**1024**	20	22	15	2
Unbalance	**8**	**2**	**6500**	8	57	2	8
A1	**20**	**2**	**3000**	14	38	19	38
A2	**35**	**2**	**5250**	18	51	33	10
A3	**50**	**2**	**7500**	59	61	47	2
S1	**15**	**2**	**5000**	12	50	15	2
S2	**15**	**2**	**5000**	40	50	13	2
S3	**15**	**2**	**5000**	32	50	15	8
S4	**15**	**2**	**5000**	40	50	20	2

Tableau III.3: Résultats des indices de validité BH, XB, WSJI, Hi**.**

Les résultats obtenus montrent l'efficacité des indices Calinski and Harabas (CH), Silhouette (SIL), Davies Bouldin (DB) et la Somme des carrés (WB) à fournir un nombre optimal de clusters par rapport aux autres indices. Ces indices prennent en compte à la fois la cohésion intra-cluster et la séparation inter-cluster. Cela permet d'évaluer à la fois la qualité de la partition des données et la capacité du clustering à séparer efficacement les différents clusters.

III.6 Conclusion

Dans ce chapitre, nous avons réalisé une comparaison sur la qualité des indices de validité en utilisant plusieurs indices internes sur différents ensembles de données de tailles et de dimensions différentes en combinant avec l'algorithme du clustering k-means. Les résultats ont montré l'efficacité des indices CH, SIL, DB et WB à obtenir le nombre optimal de clusters.

Conclusion générale

Le clustering est un processus non supervisé qui cherche à découvrir avec précision la structure inconnue des ensembles de données. Il existe de nombreuses méthodes, et l'une des plus couramment utilisées est la méthode des k-means. Le problème qui se pose est de savoir comment évaluer l'effet des algorithmes de clustering sur différents ensembles de données. La validité des clusters offre la possibilité de valider la qualité des algorithmes de clustering. Les mesures de validité des clusters sont des méthodes qui permettent de déterminer le nombre exact de clusters dans l'ensemble de données.

Dans le cadre de ce mémoire, nous proposons une comparaison consistant à comparer plusieurs indices de validité en les combinant avec l'algorithme des k-means afin de déterminer le nombre optimal de clusters. Cette méthode repose sur la variation du nombre de clusters dans une plage prédéfinie, allant de kmin à kmax, et sur l'extraction des meilleures valeurs d'indices qui représentent le nombre optimal de clusters.

Les indices de validité comparés sont les suivants : l'indice Calinski and Harabas, Hartiga, Ball and Hall, Dunn, Silhouette, Davies Bouldin, Xie Beni, WSJI, WB et le critère d'information bayésien. Ces indices sont utilisés pour évaluer la qualité des clusters formés par l'algorithme des k-means.

Afin de mener cette comparaison, nous avons réalisé des expérimentations et des comparaisons sur des ensembles de données synthétiques à différents dimensions et différents nombre de clusters.

Les résultats obtenus confirment l'efficacité de certains indices tels que les indices Calinski and Harabas, Silhouette, Davies Bouldin et la Somme des carrés (WB) dans différents ensembles de données. Ces indices se sont révélés particulièrement pertinents pour déterminer le nombre optimal de clusters.

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 le test avec d'autres indices de validité de cluster, d'autres données et d'autres algorithmes de clustering

Bibliographie

[Davis, 2006] J. Davis et M. Goadrich. The relationship between precision-recall and roccurves. Proceedings of the 23rd International Conference on Machine learning, page 233–240, 2006.

[Desgraupes, 2013] Bernard Desgraupes , Clustering Indices , article , 2013

[Far, 2021] Far balkis et Boussaadia samah, Classification des IRM cérébrales pathologiques par une approche semi-supervisée, Mémoire, 2021.

[Farhi, 2017] Hanifi Maroufel Et Habib Mahi Et Nezha Farhi , comparative study between validity indices to obtain the optimal cluster , article ,2017.

[Fawcett, 2006] T. Fawcett. An introduction to roc analysis. Pattern Recognition Letters, 27(8) : 861–874, 2006.

[Germain, 2010] Germain Forestier , Connaissances et clustering collaboratif d'objets complexes multisources , Thèse Docteur de l'Université de Strasbourg Discipline : Informatique , 2010.

[Ghribi.2011] Maha Ghribi, Pascal Cuxac, Jean-Charles Lamirel, Alain Lelu, Mesures de qualité de clustering de documents : Prise en compte de la distribution des mots clés, article, 2011.

[Gueye, 2019] NDIOUGA GUEYE, Exploration des liens formels entre les méthodes statistiques et neuronales en classification, mémoire présenté à l'université du québec à trois-rivières , 2019.

[Halkidi, 2001] Halkidi, M., Batistakis, Y., Vazirgiannis, M.: Clustering validity checking methods: Part ii (2001).

[Hennig, 2015] Christian Hennig, Recovering the number of clusters in data sets with noise features using feature rescaling factors, article, 2015.

[Jain, 1988] Jain, A.K., Dubes, R.C: Algorithms for clustering data. Prentice-Hall, Inc, Upper Saddle River, NJ, USA (1988).

[Kesavaraj, 2013] G.Kesavaraj , Dr.S.Sukumaran , A Study On Classification Techniques in Data Mining, article , 2013.

[Kharki , 2021] El Kharki, Mechbouh, Ducrot, Rouchdi, Ngono, panorama sur les methodes de classification des images satellites et techniques d'amelioration de la precision de la classification, article Article in Revue Francaise de Photogrammetrie et de Teledetection, 2021.

[Khedairia , 2014] Soufiane Khedairia , Contribution à la classification non supervisée : application aux données environnementales , these de DOCTORAT en Informatique de universite badji mokhtar annaba, 2014.

[Lallich, 2015] Stéphane Lallich, Philippe Lenca, Indices de qualité en clustering, Journée Clustering ,2015

[Laouamer ,2006] Lamri Laouamer, Approche Exploratoire Sur La Classification Appliquée Aux Images , mémoire présenté à l'université du québec à trois-rivières , 2006.

[MacQueen.1967] J.B. MacQueen , Some methods for classification and analysis of multivariate observations , Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 1967.

[Norbert, 2021] NORBERT BERTRAND SANKA, étude comparative et choix optimal du nombre de classes en classification et réseaux de neurones: application en science des données, Mémoire présenté à l'université du québec à trois-rivières , 2021.

[oubbih, 2016] OUBBIH Omar, Indice de validité en statistique décisionnelle, mémoire, 2016.

[Pakhira, 2004] Pakhira, M.K., Bandyopadhyay, S., Maulik, U.: Validity index for crisp and fuzzy clusters. Pattern Recognition 37(3), 487{501 (2004).

[Reddy , 2014] Charu C. Aggarwal Chandan K. Reddy , DATA CLUSTERING Algorithms and Applications ,Livre , 2014.

[Sanka, 2021] Norbert Bertrand Sanka, étude comparative et choix optimal du nombre de classes en classification et réseaux de neurones: application en science des données, mémoire présenté à l'université du québec à trois-rivières, 2021.

[Sublime, 2016] Jérémie Sublime, Contributions au clustering collaboratif et à ses potentielles applications en imagerie à très haute résolution, Thèse de doctorat de l'Université Paris-Saclay préparée à AgroParisTech, 2016**.**

[Tan,2005] Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining, (First Edition). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (2005).

[Zhao, 2012] Qinpei Zhao , Cluster Validity in Clustering Methods , livre , 2012.

Webographie

[S01] https://datascientest.com/algorithme-de-classification-definition-et-principauxmodeles#:~:text=Une%20classification%20supervis%C3%A9e&text=C'est%2D%C3%A0%2Ddire,pour%20ensuite%20r%C3%A9aliser%20la%20pr%C3%A9diction.

[S02] https://datascientest.com/random-forest-definition

[S03] https://www.researchgate.net/figure/Dendrogramme-de-la-classification-hierarchique-ascendante-analyse-globale_fig6_281500367/download

[S04] https://scikit-learn.org/stable/modules/mixture.html

[S05] http://abdelhamid-djeffal.net/web_documents/coursclustering1819.pdf

[S06] https://larevueia.fr/clustering-les-3-methodes-a-connaitre/

Liste des Figures

Liste des Tableaux

Résume :

Le clustering est un modèle qui explore les similitudes et les structures intrinsèques des données pour regrouper les objets en clusters, sans utiliser d'étiquettes préexistantes. Il existe de nombreuses méthodes de clustering, et l'une des plus couramment utilisées est la méthode des k-means. Dans ce mémoire, nous proposons une comparaison entre plusieurs indices de validité en les combinant avec l'algorithme des k- means afin de déterminer le nombre optimal de clusters. La méthode consiste à faire varier le nombre de clusters dans une plage prédéfinie [kmin, kmax] et extrait les meilleures valeurs d'indices représentant le nombre final de clusters. Les indices comparés sont l'indice Calinski and Harabas, Hartiga, Ball and Hall, Dunn, Silhouette, Davies Bouldin, Xie Beni, WSJI, WB et Bayesian Information Criterion. L'expérimentation et la comparaison des indices de validité ont été réalisées sur des ensembles de données synthétiques. Les résultats confirment l'efficacité de certains indices tels que les indices Calinski and Harabas, Silhouette, Davies bouldin et Somme des carrés (WB) parmi différents ensemble de données.

Les mots clé : clustering, indices de validité, K-means.

Abstract:

Clustering is a model that explores the similarities and intrinsic structures of data to group objects into clusters, without using pre-existing labels. There are many clustering methods, and one of the most commonly used is the k-means method. In this thesis, we propose a comparison between several validity indices by combining them with the k-means algorithm to determine the optimal number of clusters. The method involves varying the number of clusters within a predefined range [kmin, kmax] and extracting the best index values representing the final number of clusters. The compared indices are Calinski and Harabas index, Hartiga, Ball and Hall index, Dunn index, Silhouette index, Davies Bouldin index, Xie Beni index, WSJI, WB, and Bayesian Information Criterion. The experimentation and comparison of validity indices were performed on synthetic datasets. The results confirm the effectiveness of certain indices such as Calinski and Harabas indices, Silhouette index, Davies Bouldin index, and Sum of Squares (WB) among different datasets.

Keywords: clustering, validity indexes, K-means

ملخص :

التصنيف يتضمن تجميع البيانات المتشابهة في فئات متميزة بناء على خصائصها المشتركة. التجميع هو نموذج يستكشف التشابهات و الهياكل الجوهرية للبيانات لتجميع الكائنات في مجموعات، دون استخدام تصنيفات سابقة. هناك العديد من الطرق التجميع، ومن أحد الطرق أكثر استخداما هي طريقة k-means . في هذا البحث، نقترح مقارنة بين العديد من مؤشرات الصحة عن طريق دمجها مع خوارزمية k-means لتحديد العدد المثلى للمجموعات. تتمثل الطريقة في تغيير عدد المجموعات ضمن نطاق محدد مسبقًا [kmin, kmax] واستخراج أفضل قيم المؤشرات التي تمثل العدد النهائي للمجموعات. المؤشرات المقارنة هي مؤشر Calinski and Harabas و Hartiga و Ball and Hall و Dunn و Silhouette و Davies Bouldinو Xie Beni و WSJI و WB ومعيار المعلومات البايزية. تم إجراء التجربة ومقارنة مؤشرات الصحة على مجموعات بيانات اصطناعية. تؤكد النتائج فعالية بعض المؤشرات مثل مؤشرات Calinski and Harabas و Silhouetteو Davies Bouldin ومجموع المربعات (WB) بين مجموعات البيانات المختلفة.

كلمات المفتاحية : التصنيف، مؤشرات صحة، k-means.