ENSA
ÉCOLE NATIONALE DES SCIENCES APPLIQUÉES
KHOURIBGA

DATA MINING CLUSETERING L'algorithme OPTICS

3 mai 2023

Encadrée par

Prof : A. OURDOU

Réaliser par :
El Aattaoui Ghizlane
Mohamed Bentalb
Barj Youssef

Table des matières

1 Introduction

Le data mining, également appelé exploration de données, est un processus d'analyse de données permettant de découvrir des modèles, des relations et des tendances cachées dans les données.

Les objectifs du data mining sont multiples, notamment :

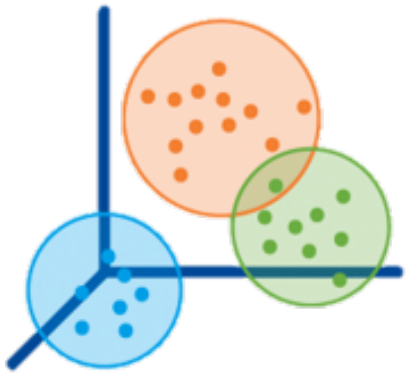
• **Prédiction** : Utilisation des modèles découverts pour prédire des événements futurs.
• **Classification** : Catégorisation des données en classes prédéfinies.
• **Clustering** : Découverte de groupes de données similaires.
• **Analyse des associations** : Identification de relations entre des variables dans les données.
Les avantages du data mining sont nombreux, tels que :

— Prise de décision : Le data mining fournit des informations précieuses pour aider les entreprises à prendre des décisions éclairées et à résoudre des problèmes complexes.
— Amélioration de la qualité des données : Le data mining peut aider à identifier et à corriger les erreurs et les données manquantes dans les bases de données.
— Connaissance du client : Le data mining permet d'analyser le comportement et les habitudes des clients pour mieux les comprendre et les cibler avec des offres et des publicités pertinentes.
— Prévention de la fraude : Le data mining peut être utilisé pour détecter les activités frauduleuses dans les transactions financières.

En résumé, le data mining est un processus important pour découvrir des modèles, des relations et des tendances cachées dans les données et peut fournir de nombreux avantages pour les entreprises et les organisations. .

2 Fondements théoriques

Dans le clustering, la mesure de similarité ou de distance est une étape importante pour identifier les groupes homogènes dans les données. La similarité mesure la ressemblance entre deux objets ou données, tandis que la distance mesure la différence ou la dissimilitude entre deux objets ou données.



Il existe plusieurs mesures de similarité et de distance utilisées dans le clustering. Voici quelques-unes des mesures les plus courantes :

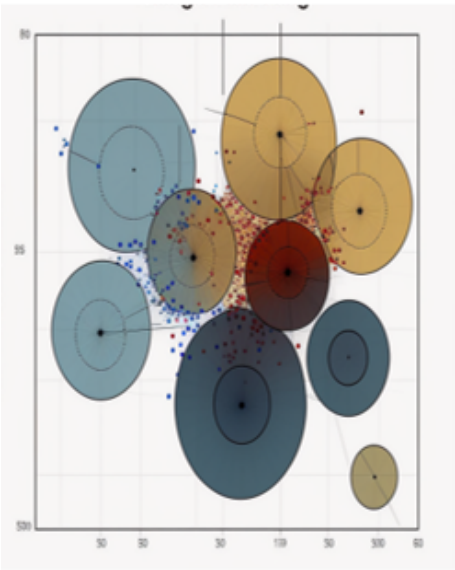
–> **Distance euclidienne** : Elle mesure la distance entre deux points dans un espace euclidien à n dimensions. Elle est largement utilisée dans le K-means et le clustering hiérarchique.

–> **Distance de Manhattan :** Elle mesure la distance entre deux points en calculant la somme des différences absolues entre les coordonnées des points. Elle est également connue sous le nom de distance de ville.

–> **Coefficient de corrélation :** Il mesure la corrélation entre deux variables. Il est souvent utilisé dans le clustering pour les données multivariées.
–> **Distance cosinus :** Elle mesure la similarité entre deux vecteurs en calculant l'angle entre eux. Elle est souvent utilisée pour des données textuelles.

3 L'Algorithme OPTICS

OPTICS (Ordering Points To Identify the Clustering Structure) est un algorithme de clustering basé sur la densité, qui peut être utilisé pour détecter des clusters de formes arbitraires et de tailles variables. Il peut également détecter les points aberrants et est plus robuste aux bruits que d'autres algorithmes de clustering.



Étapes et fonctionnement d'OPTICS

—>Calculer la distance entre chaque paire de points dans l'ensemble de données.

—>Choisir un point aléatoire et trouver ses voisins dans une certaine distance epsilon. Les points voisins sont ceux dont la distance par rapport au point initial est inférieure ou égale à epsilon.

—>Si le point initial a suffisamment de voisins (c'est-à-dire s'il y a au moins minPts points dans son voisinage), il est considéré comme un point central. Sinon, il est considéré comme un point aberrant.

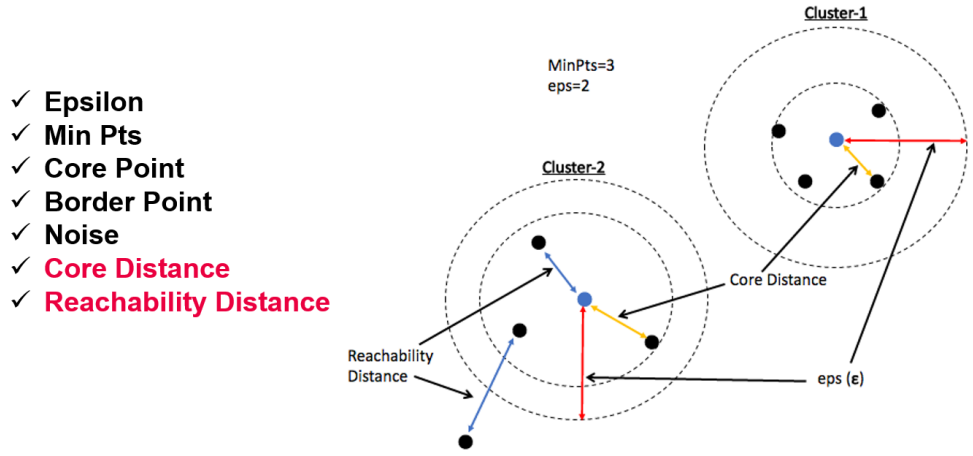
—>Pour chaque point central, calculer son rayon de portée minimum (minimum reachability distance) en fonction de la densité de ses voisins. Le rayon de portée minimum est la distance à laquelle le point central peut être atteint en passant par ses voisins les plus denses.

—> Trier les points centraux en fonction de leur rayon de portée minimum.

— > Parcourir les points centraux triés en ordre croissant de leur rayon de portée minimum. Pour chaque point central, étendre son cluster en ajoutant les points dans son voisinage qui ont une densité suffisamment élevée. Le seuil de densité est déterminé en fonction du rayon de portée minimum du point central et de la densité des points voisins. Les points qui ne sont pas dans un cluster sont considérés comme des points aberrants.

—> Répéter les étapes 4 à 6 jusqu'à ce que tous les points aient été affectés à un cluster.

fonctionnement d'OPTICS :



Utilisation d'OPTICS avec des données de densité variable :

OPTICS est particulièrement utile pour les ensembles de données avec une densité variable, car il est capable de détecter des clusters de densité variable et peut gérer les points aberrants plus efficacement que d'autres algorithmes de clustering. Dans un ensemble de données de densité variable, les clusters peuvent être de tailles et de formes variables, et certains points peuvent être isolés ou appartenir à des clusters de petite taille. L'algorithme DBSCAN, par exemple, peut avoir des difficultés à détecter les clusters de densité variable, car il utilise un seuil de densité fixe pour tous les points.

OPTICS, en revanche, utilise une mesure de densité variable pour chaque point, basée sur son rayon de portée minimum et la densité de ses voisins. Cela permet à l'algorithme de détecter des clusters de densité variable et de gérer les points aberrants plus efficacement.
OPTICS peut également être utilisé pour visualiser la structure de cluster à l'aide d'un diagramme OPTICS (un graphique qui montre les clusters et leur hiérarchie).

En utilisant OPTICS avec des données de densité variable, il est possible de découvrir des clusters de formes et de tailles arbitraires, même dans des ensembles de données avec une densité variable. Cela peut être utile dans de nombreuses applications, telles que l'analyse de données géospatiales, l'analyse de données de réseaux sociaux ou l'analyse de données scientifiques

4 Les avantages d'OPTICS

4.1 Capacité à détecter des clusters de densité variable

OPTICS est capable de détecter des clusters de tailles et de formes variables dans des ensembles de données de densité variable. Cela le rend plus robuste que d'autres algorithmes de clustering, tels que DBSCAN, qui ont des difficultés à gérer les clusters de densité variable.

4.2 Capacité à gérer les points aberrants

OPTICS est également capable de gérer les points aberrants plus efficacement que d'autres algorithmes de clustering. Il peut détecter les points aberrants et les traiter en tant que tels, plutôt que de les inclure dans des clusters où ils ne devraient pas être.

4.3 Pas de paramètres de seuil fixes

Contrairement à certains autres algorithmes de clustering, tels que DBSCAN, OPTICS ne nécessite pas de paramètres de seuil fixes. Les paramètres de seuil, tels que le seuil de densité et la distance epsilon, sont déterminés automatiquement à partir des données elles-mêmes.

4.4 Visualisation de la structure de cluster

OPTICS peut être utilisé pour générer un diagramme OPTICS, qui montre la structure hiérarchique des clusters. Cela peut être utile pour visualiser et interpréter les résultats du clustering.

4.5 Sensible à la densité de données

OPTICS est sensible à la densité de données. Les ensembles de données avec une densité très faible peuvent poser des difficultés pour OPTICS, car il peut avoir du mal à trouver des points centraux et des clusters.

5 Les inconvénient d'OPTICS

5.1 Complexité computationnelle

L'algorithme OPTICS a une complexité computationnelle relativement élevée, ce qui peut le rendre lent pour les grands ensembles de données.

5.2 Dépendance aux paramètres

L'algorithme OPTICS dépend de paramètres tels que la distance maximale entre les points et le nombre minimal de points pour former un cluster, qui doivent être ajustés pour obtenir les résultats souhaités

5.3 Limitations pour les données non-euclidiennes

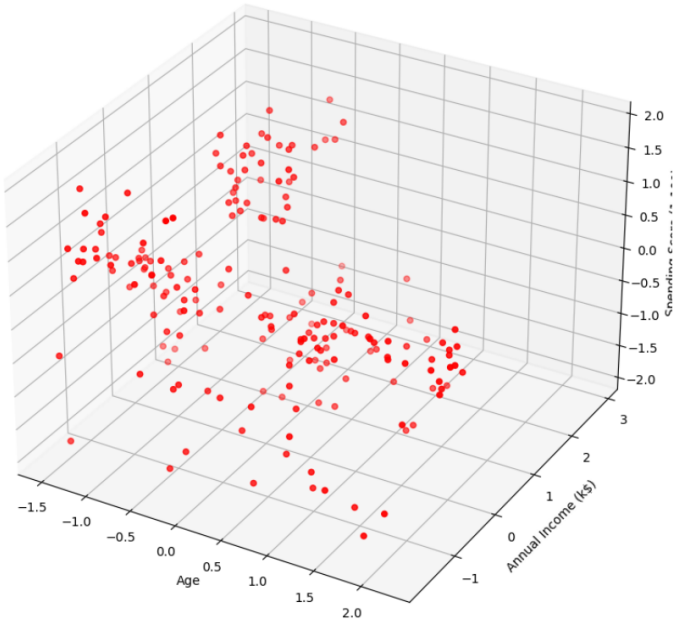
L'algorithme OPTICS est conçu pour les données euclidiennes et peut ne pas être approprié pour les données non-euclidiennes, comme les données de texte ou d'images.

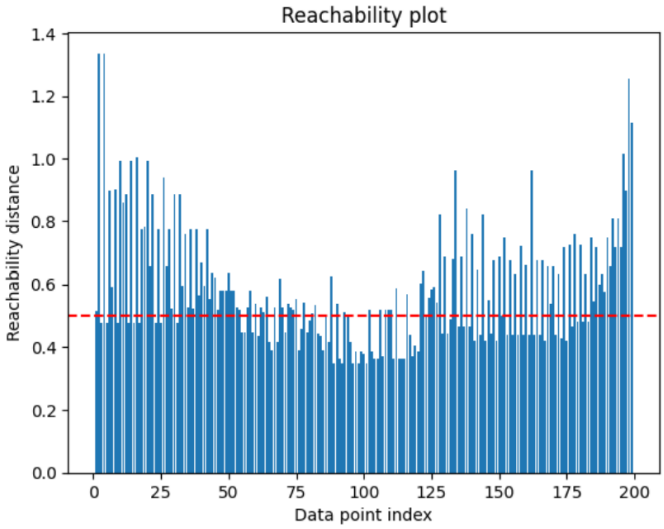
5.4 Difficulté à interpréter les résultats

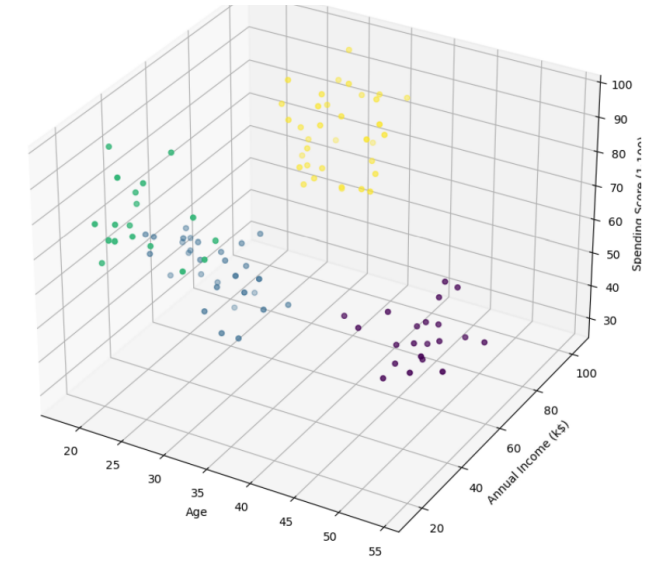
Les résultats de l'algorithme OPTICS peuvent être difficiles à interpréter et à visualiser, en particulier pour les ensembles de données de grande dimensionnalité.

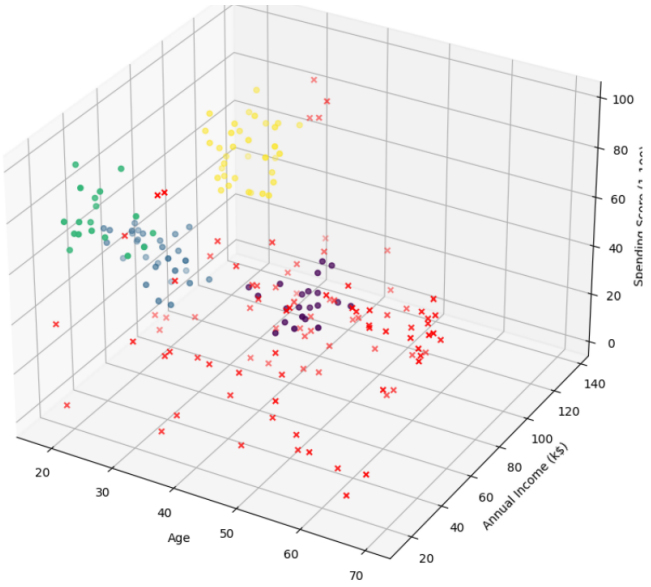
6 Exemple pratique

Application de l'algorithme OPTICS sur des données réelles






Age
Annual Income (k$)
Spending Score (1-100)


Age
Annual Income (k$)
Spending Score (1-100)

7 Les bibliothèques de traitement de données :

1. **Scikitlearn :**
   Scikit-learn est une bibliothèque open-source populaire de machine learning pour Python qui fournit des outils simples et efficaces pour l'exploration de données et l'analyse de données. Elle est construite sur NumPy, SciPy et matplotlib, et est conçue pour être facile à utiliser et bien fonctionner avec d'autres bibliothèques Python.

FIGURE 1 – Logo scikitlearn

2. **Matplotlib :**
   Matplotlib est une bibliothèque open-source de visualisation de données pour Python. Elle permet de créer des graphiques, des tableaux, des diagrammes et d'autres types de visualisations de données à partir de données en Python.

FIGURE 2 – Logo matplotlib

3. **Numpy :**
   NumPy est un outil essentiel pour les applications scientifiques et les calculs numériques en Python. Elle permet de manipuler facilement des données en grandes quantités et d'effectuer des opérations mathématiques complexes de manière efficace.

FIGURE 3 – Logo Numpy

8 conclusion

L'algorithme OPTICS est une méthode de clustering puissante qui peut être utilisée pour analyser des ensembles de données complexes avec des densités différentes. Contrairement à d'autres algorithmes de clustering, OPTICS n'exige pas de spécifier le nombre de clusters à l'avance, ce qui le rend très flexible et adaptable aux données réelles. En outre, OPTICS peut être utilisé pour identifier des clusters de forme arbitraire, contrairement à certaines autres méthodes de clustering qui ne sont efficaces que pour des clusters de formes régulières.