Résumé du Cours : Data Mining Concepts, Algorithmes et Applications

Dr. EL BENANY Mohamed Mahmoud
Assistant Professeur en Intelligence Artificielle et Science des Données

23 janvier 2025

Figure 1 – Pipeline du Data Mining

Data Mining

Data Mining

Le Data Mining (ou exploration de données) est un processus essentiel permettant de découvrir des modèles cachés, des relations significatives et des informations utiles dans de grandes bases de données. Il repose sur l'application de techniques statistiques, mathématiques et informatiques afin d'extraire des connaissances et des patterns [1]. Le processus d'extraction des connaissances à partir des données est également appelé Knowledge Discovery in Databases (KDD) [2].

Le Data Mining comprend plusieurs étapes allant de la collecte des données à l'interprétation des résultats. Ces étapes incluent la préparation des données, leur transformation, l'application de modèles d'extraction de patterns et l'interprétation des résultats [3].

A retenir :

1. **KDD et Data Mining**

2. **Pipeline du Data Mining**
3. **Méthodes de Data Mining**
   — Règles d'Association
   — Classification et Régression
   — Clustering (ou Groupement)
4. **Algorithmes**
   — Apriori (pour les règles d'association)
   — KNN (pour la classification)
   — K-Means (pour le clustering)
5. **Exercices pratiques**
6. **Interprétation**

1. KDD et Data Mining

Le processus de KDD se compose de plusieurs étapes : la collecte des données, leur prétraitement (nettoyage, transformation), l'extraction de connaissances via des méthodes de Data Mining, et enfin l'interprétation des résultats obtenus. Le but ultime est de transformer de grandes quantités de données brutes en connaissances exploitables [2].

— **Sélection des données :** Choisir les données pertinentes pour le problème à résoudre.
— **Prétraitement des données :** Nettoyer, normaliser ou transformer les données pour les rendre compatibles avec les techniques de Data Mining.
— **Transformation des données :** Réduire la dimensionnalité, appliquer des techniques de normalisation ou de réduction de bruit.
— **Extraction de connaissances :** Appliquer des algorithmes de Data Mining pour extraire des patterns cachés.
— **Interprétation des résultats :** Analyser les patterns extraits pour en tirer des informations significatives.

2. Pipeline du Data Mining

Le pipeline du Data Mining peut être divisé en plusieurs étapes clés [1] :

1. **Collecte de données :** Rassembler les données issues de différentes sources.
2. **Préparation des données :** Nettoyer et transformer les données pour les rendre prêtes à l'analyse.
3. **Transformation des données :** Appliquer des techniques comme la réduction de dimensionnalité (*e.g.*, PCA).
4. **Modélisation :** Appliquer des techniques de Data Mining pour extraire des modèles ou patterns.
5. **Évaluation des modèles :** Évaluer les performances des modèles sur des jeux de test.
6. **Interprétation des résultats :** Analyser les résultats pour en tirer des conclusions.

3. Méthodes de Data Mining

Les principales méthodes de Data Mining comprennent :

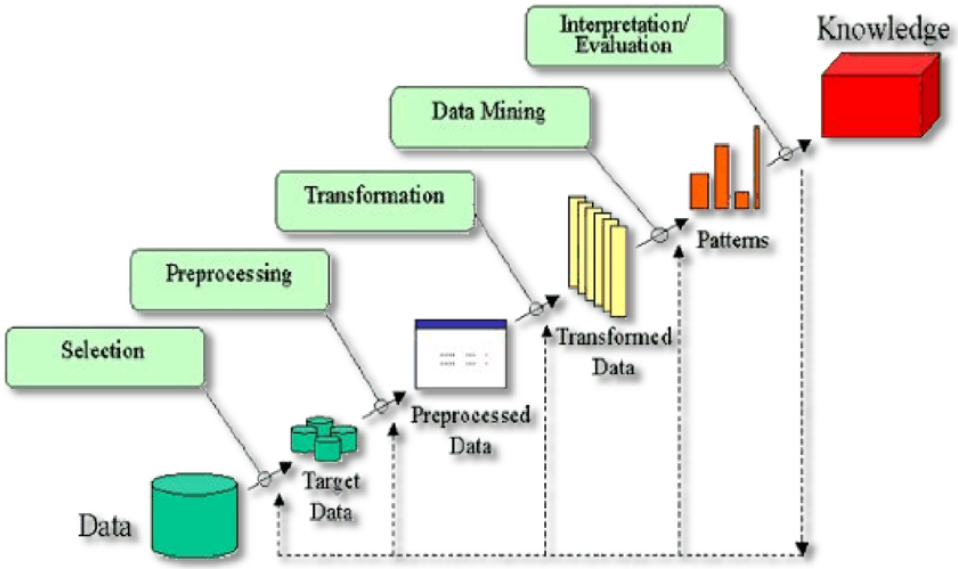

FIGURE 2 – Pipeline du Data Mining

3.1 Règles d'Association

Les règles d'association sont des règles qui indiquent des relations intéressantes entre les variables d'un ensemble de données. Par exemple, l'algorithme **Apriori**, proposé par Agrawal et Srikant [4], est couramment utilisé pour découvrir des associations dans les bases de données transactionnelles. L'algorithme Apriori permet de trouver des itemsets fréquents et d'extraire les règles d'association.

3.2 Classification et Régression

La classification prédit une étiquette (classe) pour chaque observation en fonction de ses caractéristiques. Par exemple, l'algorithme **KNN (K-Nearest Neighbors)** est une méthode populaire pour la classification supervisée [5].

3.3 Clustering

Le clustering divise un ensemble de données en groupes homogènes appelés clusters. L'algorithme **K-Means**, introduit par MacQueen [6], est couramment utilisé pour le regroupement en fonction de la similarité des données.

4. Exemples d'algorithmes

4.1 Algorithme Apriori (Règles d'Association)

L'algorithme **Apriori** est conçu pour identifier les itemsets fréquents dans les bases de données transactionnelles [4].

Étapes de l'algorithme Apriori :

1. Calculer les itemsets fréquents de taille 1.
2. Générer les itemsets de taille 2 en combinant les itemsets fréquents.
3. Répéter ce processus jusqu'à ce que les itemsets de taille k ne soient plus fréquents.
4. Générer des règles d'association à partir des itemsets fréquents.

Algorithme Apriori

ENTRÉES : Base de données de transactions D, Seuil de support minimum σ.
SORTIES : Ensemble des items fréquents.

Algorithm 1 Algorithme Apriori

$i \leftarrow 1$
$G_1 \leftarrow$ groupe des motifs de taille 1 (un seul item)
while $G_i \neq \emptyset$ **do**
    Calculer le support de chaque motif $m \in G_i$ dans la base
    $F_i \leftarrow \{m \in G_i \mid \text{support}(m) \geq \sigma\}$
    $G_{i+1} \leftarrow$ toutes les combinaisons possibles des motifs de F_i de taille $i+1$
    $i \leftarrow i+1$
end while
return $\bigcup_{i \geq 1} F_i$

4.2 Algorithme KNN (Classification)

L'algorithme **KNN** est une méthode supervisée utilisée pour la classification. Il attribue une étiquette à une observation en fonction des étiquettes des k voisins les plus proches dans l'espace des caractéristiques [5].

Étapes de l'algorithme KNN :

1. Calculer la distance entre le point à classer et tous les autres points du jeu de données.
2. Sélectionner les k voisins les plus proches.
3. Attribuer la classe majoritaire parmi les voisins.

Algorithme K-Nearest Neighbors (KNN)

ENTRÉES : Ensemble d'entraînement $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, Nombre de voisins k, Nouvelle instance x_{test}.
SORTIES : Classe prédite pour x_{test}.

Algorithm 2 Algorithme K-Nearest Neighbors (KNN)

Étape 1 : Calcul des distances
for chaque point $(x_i, y_i) \in T$ **do**
    Calculer la distance $d(x_{\text{test}}, x_i)$ (par exemple, distance euclidienne)
end for
Étape 2 : Sélection des k plus proches voisins
Trier les points (x_i, y_i) par ordre croissant de $d(x_{\text{test}}, x_i)$
Sélectionner les k premiers voisins $V_k = \{(x_{i_1}, y_{i_1}), \ldots, (x_{i_k}, y_{i_k})\}$
Étape 3 : Vote majoritaire
Identifier la classe c ayant le plus grand nombre d'occurrences parmi les voisins V_k
return Classe prédite c

4.3 Algorithme K-Means (Clustering)

L'algorithme **K-Means** est utilisé pour partitionner un ensemble de données en k clusters en minimisant la variance intra-cluster [6].

Étapes de l'algorithme K-Means :

1. Initialiser k centres de clusters aléatoires.
2. Assigner chaque point de données au cluster le plus proche.
3. Recalculer les centres des clusters.
4. Répéter les étapes précédentes jusqu'à ce que les centres des clusters ne changent plus.

Algorithme K-Means

ENTRÉES : Ensemble de données $X = \{x_1, x_2, \ldots, x_n\}$, Nombre de clusters k, Critère de convergence ϵ.
SORTIES : Ensemble des clusters $C = \{C_1, C_2, \ldots, C_k\}$ et leurs centroïdes $\mu_1, \mu_2, \ldots, \mu_k$.

Algorithm 3 Algorithme K-Means

Initialiser aléatoirement k centroïdes $\mu_1, \mu_2, \ldots, \mu_k$
repeat
  Étape d'affectation :
  for chaque point $x \in X$ **do**
    Assigner x au cluster C_j tel que $j = \arg\min_{1 \leq j \leq k} \|x - \mu_j\|^2$
  end for
  Étape de mise à jour :
  for chaque cluster C_j **do**
    Calculer le nouveau centroïde : $\mu_j \leftarrow \frac{1}{|C_j|} \sum_{x \in C_j} x$
  end for
until les centroïdes $\mu_1, \mu_2, \ldots, \mu_k$ convergent (changement $\leq \epsilon$)
return $C = \{C_1, C_2, \ldots, C_k\}$ et $\mu_1, \mu_2, \ldots, \mu_k$

5. Recapitulation

Avantages et Inconvénients de chaque algorithme

Algorithme	Avantages	Inconvénients
K-Means	Simple, rapide, scalable	Sensible aux outliers, nécessite k, clusters sphériques
KNN	Simple, non-paramétrique, flexible	Lenteur sur grandes données, sensible au choix de k
Apriori	Interprétable, puissant pour bases transactionnelles	Inefficace pour grands ensembles, sensible au seuil de support

Table 1 – Comparatif des algorithmes K-Means, KNN et Apriori.

Objectif principal de chaque algorithme

Algorithme	Objectif
K-Means	Regrouper des points de données en k clusters en minimisant la variance intra-cluster et en maximisant la distance inter-cluster.
KNN	Classer une nouvelle instance ou prédire une valeur en se basant sur la majorité ou la moyenne des k voisins les plus proches dans l'ensemble d'entraînement.
Apriori	Identifier les motifs fréquents et les règles d'association dans une base de données transactionnelle en fonction d'un seuil de support et de confiance.

TABLE 2 – Objectif principal des algorithmes K-Means, KNN et Apriori.

Relation avec le Data Mining

Algorithme	Relation avec le Data Mining
K-Means	Utilisé pour le clustering, une tâche essentielle du Data Mining permettant de regrouper les données similaires dans des clusters. Cela aide à identifier des structures ou des schémas cachés dans les données.
KNN	Appliqué dans le classification supervisée, une branche du Data Mining. KNN permet de prédire la classe ou la valeur d'une nouvelle donnée en se basant sur les données existantes.
Apriori	Utilisé pour découvrir des règles d'association, une tâche clé du Data Mining. Il identifie des relations fréquentes entre des items dans des bases transactionnelles (par exemple, l'analyse de panier d'achat).

TABLE 3 – Relation des algorithmes K-Means, KNN et Apriori avec le Data Mining.

6. Interprétation des Résultats

Une fois les modèles appliqués, il est essentiel d'interpréter les résultats. Les résultats doivent être analysés pour en tirer des conclusions pertinentes, identifier des tendances ou des anomalies, et les appliquer dans un contexte spécifique. Par exemple, dans le cas des règles d'association, l'analyse des règles générées permet de découvrir des relations cachées entre les items. Pour la classification et le clustering, l'interprétation des groupes ou des classes peut aider à comprendre les structures sous-jacentes des données.

Références

[1] Han, J., Kamber, M., & Pei, J. (2011). *Data Mining : Concepts and Techniques*. Elsevier.

[2] Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37-54.

[3] Tan, P. N., Steinbach, M., Karpatne, A., & Kumar, V. (2018). *Introduction to Data Mining*. Pearson.

[4] Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th VLDB Conference*, 487-499.

[5] Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21-27.

[6] MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281-297.