## Chapitre 2 : Méthodes de prétraitements et préparation de données

Dans ce chapitre, nous explorons la gestion de divers types de données, qu'elles soient structurées, non structurées ou semi-structurées, ainsi que la caractérisation des attributs et des mesures de distance adaptées à différentes formes de données. De plus, nous examinons les méthodes de prétraitement et de préparation des données en Data Mining. Ces étapes fondamentales sont essentielles pour assurer la qualité et la pertinence des données avant de les soumettre à une analyse plus approfondie.

### 1. Introduction aux Méthodes de Prétraitement

#### 1.1. Gestion des Données Structurées, Non Structurées et Semi-Structurées

**A. Données Structurées :**

Les données structurées sont organisées dans un format prédéfini, avec des schémas clairement définis. Elles sont généralement stockées dans des bases de données relationnelles et se prêtent bien à une analyse quantitative. Les exemples incluent les données contenues dans les feuilles de calcul Excel, les tables de bases de données SQL et les fichiers CSV. Les caractéristiques clés des données structurées sont :

- Organisation : Les données sont organisées en lignes et en colonnes, avec des relations prédéfinies entre les entités.
- Schéma Fixe : Les données suivent un schéma prédéfini, ce qui facilite la recherche et l'analyse.
- Requêtes SQL : Les données structurées peuvent être incluses à l'aide de requêtes SQL pour obtenir des informations spécifiques.

**B. Données Non Structurées :**

Contrairement aux données structurées, les données non structurées ne suivent pas de format préétabli. Elles ne s'intègrent pas facilement dans des bases de données relationnelles. Les exemples courants de données non structurées incluent le texte brut, les images, les vidéos, les fichiers audio et les documents PDF. Les caractéristiques clés des données non structurées sont :

- Manque de Schéma : Les données non structurées ne suivent pas de schéma fixe, ce qui rend leur organisation et leur analyse plus complexes.
- Diversité : Les types de données non structurées sont variés, ce qui nécessite des techniques d'analyse spécifiques à chaque type.
- Traitement Naturel du Langage (NLP) : Les techniques de NLP sont souvent utilisées pour extraire des informations significatives à partir de données textuelles non structurées.

**C. Données Semi-Structurées :**

Les données semi-structurées présentent des caractéristiques à mi-chemin entre les données structurées et non structurées. Elles ont généralement une certaine forme d'organisation, mais elles ne suivent pas strictement un schéma prédéfini comme les données structurées. Les exemples incluent les fichiers XML, les documents JSON et les pages HTML. Les caractéristiques clés des données semi-structurées sont :

- Hiérarchie : Les données semi-structurées peuvent avoir une structure hiérarchique, ce qui permet d'organiser les informations de manière plus flexible.
- Balises et Marqueurs : Les données semi-structurées utilisent souvent des balises ou des marqueurs pour définir la structure et les relations entre les éléments.
- Souplesse : Les données semi-structurées permettent une certaine souplesse dans leur organisation tout en offrant des indices sur la manière dont elles doivent être interprétées.

## 1.2. Caractérisation des Attributs

### 1.2.1. Attributs binaires :

Les attributs binaires ont deux valeurs possibles, généralement 0 et 1. Par exemple, le genre (homme, femme) est un attribut binaire.

### 1.2.2. Attributs qualitatifs et Quantitatifs

**A. Attributs qualitatifs :**

Les attributs qualitatifs, également connus sous le nom d'attributs catégoriels, représentent des caractéristiques qui ne sont pas mesurables numériquement, mais qui sont plutôt des catégories ou des étiquettes. Ils sont souvent utilisés pour décrire des propriétés non numériques d'objets ou d'entités. Les attributs qualitatifs sont fondamentaux pour diverses tâches d'analyse des données, car ils permettent de caractériser et de différencier des éléments en fonction de leurs propriétés.

Exemples :

- Couleurs (rouge, bleu, vert)
- Types de produits (alimentaire, électronique, vêtements)
- Régions géographiques (Amérique du Nord, Europe, Asie)

**B. Attributs Quantitatifs :**

Les attributs quantitatifs sont numériques et mesurables, ce qui signifie qu'ils peuvent être soumis à des opérations mathématiques telles que l'addition, la soustraction, la

multiplication, etc. Ils représentent des mesures et des quantités, ce qui en fait des éléments essentiels pour les analyses quantitatives.

Exemples :

- Âge (5 ans, 16 ans, 58 ans)
- Revenu (34.000 DA, 76.000 DA, 115.000 DA)
- Taille (1.20 M, 1.60 M, 1.82 M)

### 1.2.3. Attributs Nominaux, Ordinaux, Discrets et Continus :

#### A. Attributs Nominaux et Ordinaux :

Les attributs nominaux et ordinaux appartiennent à la catégorie des attributs qualitatifs (catégoriels) car ils sont utilisés pour décrire des caractéristiques non numériques d'objets ou d'entités.

**Attributs Nominaux :** Les attributs nominaux sont des catégories qui n'ont pas d'ordre particulier.

Exemple d'attribut nominal : Couleurs

Considérons un ensemble de voitures, chacune ayant une couleur attribuée. Les couleurs (rouge, bleu, vert, etc.) sont des catégories distinctes sans aucun ordre prédéfini. Aucune couleur n'est "plus grande" ou "plus petite" qu'une autre. C'est un attribut nominal car il représente une catégorie sans relation d'ordre.

**Attributs Ordinaux :** Les attributs ordinaux ont un ordre, mais les différences entre les valeurs ne sont pas uniformes. Par exemple, les niveaux de satisfaction (faible, moyen, élevé).

Exemple d'attribut ordinal : Niveaux de Satisfaction

Imaginons que nous ayons enquêté sur la satisfaction des clients concernant un produit et avons obtenu les niveaux de satisfaction suivants : faible, moyen, élevé. Les niveaux ont un ordre ("faible" < "moyen" < "élevé"), mais les différences entre les niveaux ne sont pas nécessairement équivalentes. C'est un attribut ordinal car il a un ordre mais pas d'échelle uniforme.

#### B. Attributs Discrets et Continus :

Les attributs discrets et continus sont des sous-catégories d'attributs quantitatifs. Ils commentent librement les valeurs sont mesurées ou quantifiées.

**Attributs discrets :** Les attributs discrets ont un ensemble fini ou dénombrable de valeurs.

Exemple d'attribut discret : Nombre d'Étudiants

Dans une classe, le nombre d'étudiants inscrits est un exemple d'attribut discret. Il ne peut prendre que des valeurs entières (0, 1, 2, ...), et il y a un nombre limité d'options possibles. Les valeurs sont distinctes et comptables.

**Attributs Continus :** Les attributs continus peuvent prendre une infinité de valeurs possibles.

Exemple d'attribut continu : Poids

Le poids d'un objet est un attribut continu. Il peut prendre une gamme infinie de valeurs possibles, et il peut inclure des valeurs décimales. Par exemple, un objet peut peser 2,5 kg, 2,501 kg, 2,50123 kg, etc. Les valeurs interrompent sont inues et ne peuvent pas être dénombrées.

### 1.3. Statistiques des données

#### A. Mesure de la tendance centrale :

Cette mesure vise à trouver une valeur représentative autour de laquelle les données se regroupent. La mesure la plus courante de la tendance centrale est la moyenne . La moyenne est exploitée en augmentant toutes les valeurs de données et en divisant le total par le nombre de valeurs.

Exemple : Supposons que nous ayons les âges de cinq personnes : 20, 25, 30, 35, et 40 ans. Pour calculer la moyenne, nous additionnons ces âges (20 + 25 + 30 + 35 + 40) et divisons par 5 (le nombre de personnes), ce qui donne une moyenne de 30 ans.

#### B. Mesure de la dispersion des données :

Cette mesure nous dit à quel point les données sont réparties autour de la tendance centrale. Une mesure courante de la dispersion est la variance. La variance mesure la moyenne des carrés des écarts par rapport à la moyenne.

Exemple : Reprenons les âges précédents. La variance serait calculée en trouvant les carrés des écarts par rapport à la moyenne, c.-à-d : $(20-30)^2$, $(25-30)^2$, $(30-30)^2$, $(35-30)^2$ , $(40-30)^2$, en faisant la moyenne de ces carrés, ce qui donnerait une valeur représentative de la dispersion.

### 1.4. Mesures de distance et de similarité

Dans le domaine du Data Mining, comprendre comment mesurer la distance et la similarité entre les données sont essentielles pour des tâches telles que le clustering, la classification et la recherche de motifs ... Ce chapitre explorera différentes mesures de distance et de similarité utilisées dans ces contextes.

**A. Distance Euclidienne**

La distance euclidienne mesure la distance "à vol d'oiseau" entre deux points dans un espace euclidien. Elle est calculée en prenant le carré racine de la somme des carrés des différences entre les coordonnées des points.

Formule :

Distance (x, y) = $\sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$

Exemple :

Prenons deux points x (2, 3) et y (5, 7). La distance euclidienne entre ces deux points serait :

Distance (x, y) =$\sqrt{(2-5)^2 + (3-7)^2} = \sqrt{9+16} = \sqrt{25} = 5$

**B. Distance de Manhattan**

La distance de Manhattan, également appelée distance de la ville, mesure la distance entre deux points en utilisant la somme des distances absolues entre leurs coordonnées.

Formule :

Distance (x, y) = $\sum_{i=1}^{n} |X_i - Y_i|$

Exemple :

Reprenons les points x (2, 3) et y (5, 7). La distance de Manhattan entre ces deux points serait :

Distance (x, y) =|2−5|+|3−7|=3+4=7

**C. Distance de Minkowski**

La distance de Minkowski est une généralisation des distances Euclidienne et de Manhattan. Elle prend un paramètre P qui détermine le type de distance (quand P=2, elle est équivalente à la distance Euclidienne ; quand P=1, elle est équivalente à la distance de Manhattan).

Formule :

Distance (x, y) = $\sqrt[P]{\sum_{i=1}^{n} |X_i - Y_i|^P}$

Exemple :

Verser P=3, reprenons les points x (2, 3) et y (5, 7). La distance de Minkowski serait :

Distance (x, y) =$\sqrt[3]{|2-5|^3+|3-7|^3} = (27+64)^{1/3} = 91^{1/3} \approx 4.4979$

**D. Distance des données binaires**

Soit la table de contingence (dissimilarité) :

| | Objet y | | | |
|---|---|---|---|---|
| Objet x | | 1 | 0 | Sum |
| | 1 | A | B | A+B |
| | 0 | C | D | C+D |
| | Sum | A+C | B+D | P |

**Coefficient de Correspondance Simple**

Formule :

Distance (x, y) = $\frac{B+C}{A+B+C+D}$

Exemple :

X= (0 , 1 , 1 , 0 , 0 , 0 , 1 , 0 , 0 , 1)

Y=( 1 , 0 , 0 , 1 , 0 , 1 , 0 , 0 , 1 , 1)

| | Objet y | | | |
|---|---|---|---|---|
| Objet x | | 1 | 0 | Sum |
| | 1 | 1 | 3 | 4 |
| | 0 | 4 | 2 | 6 |
| | Sum | 5 | 5 | 10 |

Distance (x, y) =$\frac{3+4}{1+3+4+2} = \frac{7}{10} = 0.7$

**Coefficient de Jaccard**

Formule :

Distance (x, y) = $\frac{B+C}{A+B+C}$

Exemple :

X= (0 , 1 , 1 , 0 , 0 , 0 , 1 , 0 , 0 , 1)

Y=( 1 , 0 , 0 , 1 , 0 , 1 , 0 , 0 , 1 , 1)

| | | Objet y | | |
|---|---|---|---|---|
| Objet x | | 1 | 0 | Sum |
| | 1 | 1 | 3 | 4 |
| | 0 | 4 | 2 | 6 |
| | Sum | 5 | 5 | 10 |

Distance (x, y) =$\frac{3+4}{1+3+4} = \frac{7}{8} = 0.875$

**E. Distance pour les Données Mixtes**

Lorsque les données contiennent à la fois des attributs numériques et catégoriels, une mesure de distance appropriée est nécessaire. Une approche courante consiste à utiliser une combinaison de différentes distances pour mesurer la similarité entre les différents types d'attributs. Par exemple, vous pourriez opter pour l'utilisation de la distance euclidienne pour les attributs numériques, tandis que le coefficient de correspondance simple pourrait être employé pour les attributs catégoriels. En combinant ces mesures, il devient possible d'obtenir une mesure de distance globale qui prend en compte à la fois les aspects numériques et catégories des données mixtes.

**2. Normalisation et Standardisation :**

Ces deux étapes sont essentielles dans le prétraitement des données, en particulier lorsqu'on travaille avec des caractéristiques (attributs) qui ont des échelles différentes. Le but est de rendre les données comparables et de garantir que les variations dans une caractéristique ne dominent pas les analyses par rapport à d'autres caractéristiques.

**Normalisation :** Cette technique ajuste les valeurs d'une caractéristique pour qu'elles tombent dans une plage commune, généralement entre 0 et 1. Cela permet de mettre en évidence les relations relatives entre les valeurs, tout en préservant les structures de distribution originales. La normalisation est utile lorsque la distribution des données n'est pas nécessairement gaussienne et que les valeurs minimales et maximales peuvent varier considérablement. Par exemple, si vous avez des données de tailles d'animaux allant de 50 cm à 500 cm, la normalisation rééchelonnera ces valeurs pour qu'elles soient comprises entre 0 et 1 en fonction de leur position dans cette plage.

Exemple : Supposons que vous ayez des données de taille d'animaux :

Animal A : 50 cm

Animal B : 150 cm

Animal C : 500 cm

Après normalisation, les valeurs pourraient ressembler à :

Animal A : 0,0

Animal B : 0,5

Animal C : 1,0

**Standardisation :** Contrairement à la normalisation, la standardisation transforme les valeurs d'une caractéristique de telle sorte qu'elles présentent une moyenne de 0 et un écart-type de 1. Cela est particulièrement utile lorsque les données suivent une distribution normale. La standardisation est plus adaptée lorsque les valeurs varient considérablement en termes d'écart-type. Par exemple, si vous avez des données de températures en degrés Celsius et Fahrenheit, la standardisation ajustera ces valeurs pour qu'elles soient centrées autour de zéro et exprimées en termes d'écart-type.

Exemple : Supposons que vous ayez des données de température :

Jour 1 : 20°C (68°F)

Jour 2 : 30°C (86°F)

Jour 3 : 10°C (50°F)

Après standardisation, les valeurs pourraient être modifiées pour avoir une moyenne de 0 et un écart-type de 1 en termes d'échelle z :

Jour 1 : -0,53 z

Jour 2 : 1,06 z

Jour 3 : -0,53 z

**3. Gestion des Données Manquantes :**

Elle consiste à traiter les valeurs manquantes dans les enregistrements afin d'éviter tout biais ou inexactitude dans les résultats.

Plusieurs approches sont utilisées pour gérer les données manquantes :

**Suppression des Lignes :** Dans cette approche, les enregistrements contenant des données manquantes sont simplement supprimés. Cela peut être efficace, mais cela peut réduire la taille de l'ensemble de données, ce qui peut potentiellement affecter la représentativité des données restantes.

Exemple : Dans un sondage sur les préférences alimentaires, si certains participants n'ont pas répondu à la question sur leur plat préféré, vous pourriez choisir de supprimer ces

participants de l'analyse pour obtenir un ensemble de données complet pour cette question spécifique.

**Imputation des Valeurs :** L'imputation consiste à estimer les valeurs manquantes en se basant sur les données disponibles. Une méthode courante est de remplacer les valeurs manquantes par la moyenne, la médiane ou d'autres statistiques des valeurs similaires dans le même groupe ou contexte. Cette approche permet de conserver la taille de l'ensemble de données tout en évitant de perdre des informations.

Exemple : Si vous analysez les données de ventes mensuelles et qu'il manque certaines valeurs pour un mois, vous pourriez remplir ces valeurs manquantes en utilisant la moyenne des ventes des mois environnants pour éviter les distorsions dans l'analyse.

**4. Réduction de Dimension :**

La réduction de dimension est une technique utilisée dans l'analyse de données pour simplifier les ensembles d'attributs complexes tout en préservant autant que possible les informations importantes. L'Analyse en Composantes Principales (PCA) est une méthode fréquemment utilisée pour atteindre cet objectif.

Cette méthode transforme un ensemble d'attributs interdépendants en un nouveau jeu de dimensions non corrélées appelées composantes principales. Les composantes principales sont triées par ordre d'importance en termes de variance qu'elles capturent. Cela signifie que les premières composantes principales expliquent la majeure partie de la variabilité des données, tandis que les suivantes expliquent de moins en moins.

Exemple 1 : Si nous avons un grand nombre d'attributs décrivant les caractéristiques d'un produit, PCA peut nous aider à identifier les attributs les plus importants qui contribuent le plus à la variabilité des données.

Exemple 2 : Si nous avons des données de caractéristiques d'images, PCA peut nous aider à identifier les combinaisons linéaires d'attributs qui expliquent le plus de variance dans les données.

**5. Transformation des Attributs :**

La transformation des attributs consiste à créer de nouvelles caractéristiques pour en créer de nouveaux qui capturent mieux les informations à partir des attributs existants. Cela peut inclure des opérations mathématiques, des extractions de motifs ou des conversions de formats.

Exemple 1 : Si nous avons une base de données contenant des dates de naissance, nous pouvons créer une nouvelle caractéristique d'âge en soustrayant la date de naissance de la date actuelle.

Exemple 2 : En combinant les attributs de largeur, hauteur et longueur d'un objet, nous pouvons créer une nouvelle caractéristique qui représente le volume de l'objet.

**6. Détection d'Outliers :**

La détection d'outliers consiste à identifier les valeurs aberrantes qui diffèrent considérablement du reste des données. Ces valeurs peuvent fausser les analyses, et il est important de les identifier et de décider si elles doivent être traitées ou supprimées.

Exemple : Dans un ensemble de données sur les revenus, un revenu extrêmement élevé ou faible par rapport à la moyenne peut être considéré comme un cas aberrant.

**Conclusion**

En conclusion, ce chapitre sur les méthodes de prétraitements et la préparation de données constitue une base solide pour explorer le monde complexe de l'analyse de données et du Data Mining. Nous avons parcouru les différentes formes de données - structurées, non structurées et semi-structurées - en mettant en lumière leurs caractéristiques distinctes. La caractérisation des attributs, qu'ils soient qualitatifs ou quantitatifs, a été présentée comme une étape cruciale pour comprendre et différencier les données en vue d'une analyse approfondie.

La compréhension des mesures de distance et de similarité, telles que la distance euclidienne, la distance de Manhattan et d'autres, est essentielle pour des tâches clés telles que le clustering, la classification et la recherche de motifs. Ces méthodes nous permettent de quantifier la similarité entre des points de données, facilitant ainsi la découverte de modèles significatifs.

Nous avons également exploré des concepts essentiels tels que la gestion des données manquantes, la réduction de dimension et la transformation des attributs. Ces processus visent à optimiser la qualité et la pertinence des données, permettant ainsi des analyses plus précises et des résultats plus fiables.