2021/2022

Résumé du livre

Data Mining. Concepts and Techniques, 3rd Edition

Chapitre 1 : Introduction

1 .1 Pourquoi l'exploration de données?

L'informatisation de notre société et le développement rapide de puissants outils de collecte et de stockage de données à résulter des téraoctets ou pétaoctets de données versées dans nos réseaux informatiques qui préviennent des entreprises, société, sciences et ingénierie, médecine et presque tous les autre domaine de la vie quotidienne. Cette une énorme quantité de données doit être analysée, et a conduit à la naissance de l'exploration de données.
Depuis les années 1960, la technologie de l'information et le secteur de la gestion de bases de données (la collecte de données, création de bases de données, la gestion et analyse avancée des données) ont évolué de manière systématique, cela a conduit à un grand nombre d'ordinateurs puissants et abordables, d'équipements de collecte de données et supports de stockage. Cette technologie donne un coup de fouet à la base de données et permet à un très grand nombre de bases de données et de référentiels d'informations d'être disponible pour la gestion des transactions, la recherche d'informations et l'analyse de données.

1.2 Qu'est-ce que l'exploration de données?

L'exploration de données est définit est l'exploration de connaissances à partir de données, on peut présenter ce processus de découverte des connaissances par une séquence itérative des étapes suivantes :

Etape 1 : prétraitement des données

1. Nettoyage des données (pour éliminer le bruit et les données incohérentes)
2. Intégration des données (lorsque plusieurs sources de données peuvent être combinées)
3. Sélection de données (où les données pertinentes pour la tâche d'analyse sont extraites de la base de données)
4. Transformation de données (où les données sont transformées et consolidées dans des formulaires approprié pour l'extraction en effectuant des opérations de résumé ou d'agrégation)

Etape 2 : Exploitation de données

5. Data mining (processus essentiel où des méthodes intelligentes sont appliquées pour extraire modèles de données)
6. Évaluation des modèles (pour identifier les modèles vraiment intéressants représentant la connaissance basée sur des mesures d'intéressement)
7. Présentation des connaissances (où techniques de visualisation et de représentation des connaissances sont utilisés pour présenter les connaissances extraites aux utilisateurs)

1.3 Quels types de données peuvent être extraites ?

L'exploration de données peut s'appliquer à différentes formes de données à condition que le les données sont significatives pour une application cible. Les formes les plus élémentaires de données pour l'exploitation sont :

1.3.1 Données de base de données :

Consiste en une collecte de données interdépendantes, appelée base de données et un ensemble de logiciels qui permettent la gestion de données (définition de structure de base de données, stockage de données, spécification et la gestion simultanée, partagée), l'accès distribué aux données, assurer la cohérence et la sécurité des informations …

1.3.2 Entrepôts de données :

Un entrepôt de données est un référentiel des informations collectées à partir de sources multiples dont les données sont organisées sur des sujets principaux (par exemple, client, article, fournisseur et activité) pour faciliter la prise de décision, stockées sous une base de données unifiée. Avec un schéma sur un site unique.

1.3.3 Données transactionnelles

En général, chaque enregistrement d'une base de données transactionnelle capture une transaction, telle qu'un client, une réservation de vol ou les clics d'un utilisateur sur une page Web. Une transaction typiquement comprend un numéro d'identification de transaction unique et une liste des éléments constituant la transaction

1.3.4 Autres types de données :

Consiste en beaucoup d'autres données que Différents types de connaissances peuvent être extraites de ces types de données, ainsi, on peut citer par exemple les données séquentielles (comme les enregistrements historiques, données boursières, séries chronologiques et données biologiques), données de séquence, les flux de données (les données de surveillance de vidéo et de capteur), données spatiales (par exemple, cartes), données de conception technique (par exemple, conception de bâtiments, de composants de système ou de circuits intégrés)…

1.4 Les types de motifs qui peuvent être extraits :

Pour déterminer le type de model à trouver dans les taches (descriptive et prédictive) de l'exploration de données, il existe un certain nombre de fonctionnalités d'exploration de données :

1.4 La caractérisation et discrimination :

La caractérisation est un résumé des caractéristiques générales ou caractéristiques d'une classe de données cible et La discrimination de données est une comparaison des caractéristiques générales des objets de données de la classe cible avec les caractéristiques générales des objets d'une ou de plusieurs classes contrastantes.

1.4 .1 L'exploitation de modèles fréquents, d'associations et corrélations :

Les modèles fréquents, sont des modèles qui apparaissent fréquemment dans les données. Il existe de nombreux types de modèles fréquents, notamment des ensembles d'éléments fréquents, des modèles séquentiels et sous-structures fréquentes.

1.4 .2 Classification et régression :

La classification est le processus de recherche d'un modèle qui classe les données ou les concepts celons des classe prédéfini (connues) , la régression fait de même mais on connait pas les étiquettes de classes.

1.4 .3Analyse de regroupement (clustering):

Utilisé pour générer une étiquette de classe pour un groupe de data, le regroupement analyse les objets de données sans consulter les étiquettes de classe. Dans de nombreux cas, étiquetés les données peuvent tout simplement ne pas exister au début.

1.4 .4 Analyse des valeurs aberrantes :

Appelée extraction d'anomalie. Dans ce type d'analyse on s'intéresse à la détection des valeurs aberrantes comme le bruit et les exceptions qui peuvent être détectées à l'aide de tests statistiques.

1.5 Les technologies utilisées dans l'exploration de données

L'exploration de données a incorporé de nombreuses techniques d'autres domaines tels que les statistiques, l'apprentissage automatique, la reconnaissance de formes, la base de données et entrepôts de données, récupération d'informations, visualisation, algorithmes, haute performance informatique et de nombreux domaines d'application

1.6 Quels types d'applications sont ciblés?

L'exploration de données à jouer un rôle clé et critique dans un grand nombre d'applications tels que la bioinformatique et le génie logiciel nous discutons brièvement deux exemples

d'exploitation de données très réussis et populaires: intelligence économique et les moteurs de recherche.

1.6.1 Intelligence économique :

Il est primordiale que les entreprises soit en mesure d'analyser efficacement le marché, comparer les commentaires des clients sur des produits, découvrir les forces et les faiblesses de leurs concurrents, conserver une haute clients précieux et prendre des décisions d'affaires avisées,, cela se fait par l'utilisation de l'exploration de données

1.6.2 Moteurs de recherche Web :

Les moteurs de recherche recherchent et renvoient également des données disponibles dans des bases de données publiques ou des annuaires ouverts, et sont essentiellement de très grandes applications d'exploration de données. Diverses données

Les techniques d'exploitation minière sont utilisées dans tous les aspects des moteurs de recherche

Chapitre2 : Apprendre à connaître vos données

2.1 Les données et les types d'attributs :

Avant de passer au data mining, il est nécessaire de préparer les données (le prétraitement de données) pour le faire il faut savoir répondre a les questions suivantes : quels sont les types d'attributs ou de champs qui constituent vos données ? Quel genre de valeurs fait chaque attribut avoir ? Quels attributs sont discrets et lesquels ont une valeur continue ? Comment sont distribuées les valeurs ? Existe-t-il des moyens de visualiser les données pour avoir une meilleure idée de tout cela ?

Un attribut est un champ de donne qui représente une caractéristique d'un objet, un objet étudiant par exemple peut être décrit par les attributs : ID, nom, prénom… ect.

L'ensemble des attributs d'un objet de donnée est appelé un vecteur d'attributs.

Il existe plusieurs types d'attributs :

- Attributs nominaux : Les valeurs sont des symboles (des noms)
  Exemple :
  Les valeurs de Temps sont {Ensoleillé, Pluvieux, Neigeux, Gris}

- Attributs binaires : c'est un attribut nominal qui peux prendre que deux valeurs le 0 et le 1, généralement le 0 pour l'absence da attribut et le 1 pour sa présence.

- Attributs ordinaux : Une notion d'ordre s'impose sur les attributs ordinaux
  Mais il n'est pas possible de calculer directement des distances entre des valeurs ordinales
  Exemple :

La température est décrite par les adjectifs {chaud, froid, moyen}, et chaud > moyen > froid

- Attributs numériques : sont des attributs quantitatif, c'est à dire il s'agit d'une entité mesurable représenté dans des valeurs entières ou réelles.
  Cet attribut peut être de type intervalle ou de type rapport (ratio).

- Attributs continu ou discret : tous les attributs que nous avons représentés sont soit discrets soit continus.

  Les attributs discrets : un ensemble infini de valeurs finies.

2.2 Description statistique de donnée :

Des descriptions statistiques peuvent être utilisées pour identifier les propriétés des données et mettre en évidence les valeurs de données à traiter comme le bruit, les valeurs manquantes, les valeurs aberrantes.

Nous allons traiter trois domaines de description :

1. Nous commençons par les mesures centrales, il existe plusieurs manières pour mesurer la tendance centrale des données :

supposons qu'on a un attribut x salaire pour un ensemble d'objets (x1 ,x2 ,… ,xn). Comment savoir ou tomberait la plupart des valeurs de ce salaire ? Cela donne l'idée a la tendance centrale qui peut être calcule par la moyenne, la médiane ou le mode.

La moyenne $\bar{x}$:

$$\bar{x} = \frac{\sum_{i=1}^{i=n} x_i}{N}$$

La médiane : C'est la valeur moyenne dans un ensemble de valeurs de données ordonnées. C'est la valeur qui sépare la moitié supérieure d'un ensemble de données de la moitié inférieure.

En probabilité et en statistique, la médiane s'applique généralement aux données numériques. Supposons qu'un ensemble de données donné de N valeurs pour un attribut X est trié par ordre croissant. Si N est impair, alors la médiane est la valeur moyenne de l'ensemble ordonné. Si N est pair, alors la médiane n'est pas unique ; ce sont les deux valeurs du milieu.

Le mode : C'est la données qui se repete le plus

Exemple age : 15 20 22 22 22 24 35 35 35 38 39 42

Une variable age est donnée ainsi

Le mode ici est 35 et cette variable est bimodale car il y a deux valeurs qui se repetent 3 fois

Lorsque la variable possede un seul mode et que ses valeurs verifient :

$$moyenne - mode = 3*(moyenne-mediane)$$

alors on dit qu'elle est symetrique

2. Afin de savoir comment les données sont reparties nous passons à la plage, boite à moustache la variante et l'écart type.

La plage : la plage de l'ensemble est la différence entre la plus grande valeur (max ()) et la plus petite valeur (min ())

La variance :

$$\sigma^2 = \frac{1}{N} \sum_{I=1}^{i=N} (x_i - \bar{x})^2$$

Les quartiles : ce sont les trois points qui divisent l'ensemble de données en quatre parties égales

Chaque partie représenté un quart de la distribution.

3. Enfin, nous allons traiter quelques affichages graphiques pour les données
Il existe plusieurs affichages graphiques pour la représentation de données, comme le diagramme a barre, camembert et les histogrammes.

Histogramme : Le tracé des histogrammes est une méthode graphique permettant de résumer la distribution d'un attribut donne x.

2.3 Visualisation de données :

Afin de transmettre les données aux utilisateurs, nous utilisons la visualisation de données. Cette dernière est utilisée dans plusieurs applications.
Il existe plusieurs techniques de visualisation de données, on cite :

- *Techniques de visualisation orientées pixel :*

Un moyen simple de visualiser la valeur d'une dimension consiste à utiliser un pixel où la couleur du pixel reflète la valeur de la dimension.

- *Techniques de visualisation par projection géométrique :*

Cette technique permet de comprendre la distribution des données dans un espace multidimensionnel

- *Techniques de visualisation basées sur des icônes :*

Les techniques de visualisation à base d'icônes utilisent de petites icônes pour représenter des images multidimensionnelles

- *Techniques de visualisation hiérarchique :*

Techniques de visualisation hiérarchique partitionne toutes les dimensions en sous-ensembles (c.-à-d. sous-espaces). Les sous-espaces sont visualisés dans une manière hiérarchique.

2.4 Mesure de la similarité et de la disparité des données :

Dans le data_mining pour faire la classification, le regroupement (clustring) et d'autres explorations de données nous avons besoin de connaitre la ressemblance entre nos objets, et ça revient à calculer la similarité et la disparité entre deux objets.

Les matrices :

Supposons que nous ayons n objets (par exemple, personnes, éléments ou cours) décrits par p attributs telles que (l'âge, la taille, le poids ou le sexe).Les objets sont : x1={x11, x12,…, x1p } ; x2={x21, x22,…, x2p} etc., où xij est la valeur de l'objet xi pour le j eme attribut.

La disparité matricielle :

$$\begin{bmatrix} 0 & \cdots & d(1,3) \\ \vdots & \ddots & \vdots \\ d(3,1) & \cdots & 0 \end{bmatrix}$$

La distance entre l'objet et lui-même est0
La distance entre l'objet 1 et 3 est d(1,3)

La similarité et la disparité entre deux attributs nominaux :

La distance entre deux attributs nominaux i et j se calcule par la forme suivante :

$$d(i,j) = \frac{p-m}{p}$$

Ou m est égal au nombre des attributs entre i et j, et p le nombre total des attributs décrivant les objets.

La similarité :

$$Sim(i,j) = 1 - d(i,j) = \frac{m}{p}$$

La similarité et la disparité entre deux attributs binaires :

L'attribut binaire accepte seulement deux valeurs 1 et 0, on calcule la distance entre deux attributs de ce type comme suit ***:***

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

Ou q est égal au nombre d'attributs qui prennent la valeur 1 pour les deux objets i et j,

r égal au nombre d'attributs qui prennent la valeur 1 pour l'objet i et la valeur 0 pour l'objet j,
s égal au nombre d'attributs qui prennent la valeur 0 pour l'objet i et la valeur 1 pour l'objet j,
et t égal au nombre d'attributs qui prennent la valeur 0 pour les deux objets i et j.
p est le nombre total des attributs,

$$p=r+s+t+q$$

La similarité :

$$sim(i,j)=1-d(i,j)=\frac{r+s}{q+r+s+t}$$

La similarité et la disparité entre deux attributs numériques :
La distance entre deux attributs numériques, elle peut être calculé par différentes méthodes, c'est la distance entre deux points, on peut utiliser la distance Euclidienne ou Manhattan …ect

$$d(i,j)=\sqrt{(x_{i1}-x_{i1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{ip}-x_{jp})^2}$$

Chapitre 3 : Prétraitement des données

3.1 Prétraitement des données :

Pourquoi prétraiter les données?

Les données sont de qualité si elles répondent aux exigences de l'utilisation prévue, telle que l'exactitude, l'exhaustivité, la cohérence, l'actualité, crédibilité et interprétabilité.

3.2 Nettoyage des données.

Processus de nettoyage de données :

3.2.1 Valeurs manquantes :

1. Ignorer le tuple
2. Remplir la valeur manquante manuellement
3. Utiliser une constante globale pour renseigner la valeur manquante
4. Utiliser une mesure de la tendance centrale de l'attribut (par exemple, la moyenne ou la médiane) pour remplir la valeur manquante
5. Utiliser l'attribut moyen ou médian pour tous les échantillons appartenant à la même classe que le tuple donné:
6. Utiliser la valeur la plus probable pour renseigner la valeur manquante soit le mode quand c'est une variable nominale
   Quand c'est une variable numerique on peut faire une tache d'estimation également

3.2.2 Données bruitées

Le bruit est une erreur ou une variance aléatoire dans une variable mesurée. Parmis les méthodes qui élimine le bruit on peut citer :

- Binning: les méthodes de binning lissent une valeur de données triée en consultant son "voisinage" c'est-à-dire les valeurs qui l'entourent.
- Régression: une technique conforme aux valeurs de données à une fonction. La régression linéaire consiste à trouver la «meilleure» ligne à adapter deux attributs (ou variables) afin qu'un attribut puisse être utilisé pour prédire l'autre.

- Analyse des valeurs aberrantes: les valeurs aberrantes peuvent être détectées par regroupement, par exemple, lorsque les valeurs sont organisées en groupes ou «grappes».

3.2.3 Le nettoyage des données en tant que processus :

La première étape du nettoyage des données en tant que processus est la détection des incohérences, les données doivent également être examinées en ce qui concerne les règles uniques, les règles consécutives et les valeurs nulles…

3.3 Intégration des données :

Il y a un certain nombre de problèmes qui doivent être pris en compte lors de l'intégration de données :

- Problème d'identification d'entité
- Analyse de redondance et de corrélation
- Duplication de tuple
- Détection et résolution de conflits de valeurs de données

3.4 Réduction des données :

Des techniques de réduction des données peuvent être appliquées pour obtenir une représentation réduite de l'ensemble de données beaucoup plus petit en volume, tout en maintenant de près l'intégrité du document d'origine.

- Aperçu des stratégies de réduction des données
- Transformée en ondelettes
- Analyse en composantes principales
- Sélection de sous-attributs
- Modèles de régression et log-linéaires
- Les histogrammes
- Clustering
- Échantillonnage
- Agrégation de cube de données

3.5 Transformation des données et discrétisation des données :

Les stratégies de transformation de données :

- Le lissage
- Construction d'attribut
- Agrégation
- Normalisation
- Discrétisation
- Génération de hiérarchie de concepts pour les données nominales