Université des Sciences et de la Technologie Houari Boumediene

Data Mining

Résumé du livre "Data Mining Concepts and Techniques" Chapitres 1, 2 et 3

Rédaction:
MOULAI Hassina Safaa
Matricule : 201400007564
HOUACINE Naila Aziza
Matricule : 201400007594
M2 SII Groupe:3

Professeur
Mme. BABA ALI

October 6, 2018

Contents

Chapter 1

Introduction[1]

l'apparition est du à la richesse du monde actuel en données et l'abondance de ces derniers mais paradoxalement pauvre en information , la croissance exponentielle en quantité de données qui sont stockées et collectées sont à l'origine de l'incapacité humaine à gérer ce flux de données et en extraire des connaissances pertinentes sans une utilisation d'outils adéquates, le besoin en exploration de données grandit et.

1.1 Qu'est ce que le Data mining

Le Data mining ou fouille de données est le processus d'exploration et d'extraction de connaissances ou de motifs à partir d'un volume conséquent de données en utilisant des outils , le processus comporte plusieurs étapes qui sont les suivantes :

1.1.1 Nettoyage de données

ceci consiste à éliminer les données bruitées

1.1.2 Integration de données

1.1.3 Sélection de données

on récupéré les données pertinentes de la base de données pour qu'elles puissent par la suite être utilisé lors de l'analyse

1.1.4 Transformation de données

La transformation consiste à faire des opérations d'agrégation , normalisation ou bien un résumé à fin de rendre les données sous une formes approprié pour l'exploration minière .

1.1.5 Data mining

c'est l'application de méthodes intelligentes pour l'extraction des modèles ou motifs de données

1.1.6 Évaluation des motifs

c'est une Identification des motifs pertinents représentant le mieux la connaissance

1.1.7 Présentation de données

c'est la visualisation et présentation des connaissances aux utilisateurs.

1.2 Les types de données qui peuvent être utilisé pour le data mining

Ils existent différent types de données sur les quelles le data mining peut opérer , pour cella la seule condition suffisante est que ces données doivent avoir du sens , parmi elles :

1.2.1 Les bases de données

c'est une collection de données interdépendantes un logiciel (système de gestion de base de données SGBD) permettant de gérer ,accéder et sécuriser ces données, les données sont sous forme de tables relationnelles suivant un certain format, chaque table comporte des colonnes qui appelées attribues et les lignes sont les tuples ou instance (un objet).

1.2.2 Les entrepôts de données

un entrepôt de données est un référentiel des informations collectées à partir de sources multiples, stockées sous une base de données unifiée. ce schéma est résidant généralement sur un site unique, l'entrepôt de données est construit via un processus de nettoyage des données, d'intégration, de transformation, de chargement et de rafraîchissement des données odiques et généralement modélisé par une structure de données multidimensionnelle, appelée cube de données dans lequel chaque dimension représente un ou plusieurs attributs ,ce cube de données fournit une vue multidimensionnelle des données et permet le précalcul et l'accès rapide aux données résumées.

1.2.3 Base données transactionnelles

une base de données transactionnelles se compose principalement de transactions mais peut aussi contenir des tables supplémentaires contenant des informations sur les transactions existantes .

1.2.4 Autres types de données

il existe d'autres types de données sous formes et structures polyvalentes et assez différentes sémantiquement. Ces types de données peuvent être vus dans de nombreuses applications: liées au temps données séquentielles, les flux de données , données spatiales (cartes, par exemple), données de conception technique , hypertexte et multi-données multimédia (y compris texte, image, vidéo et audio), graphique et données en réseau et le Web . Ces applications apportent de nouveaux défis, tel que la gestion des données contenant des structures spéciales et une sémantique spécifique

1.3 Quelles sont les motifs qui peuvent être extraites ?

1.3.1 Classe/Concept Description: Caractérisation et Discrimination

Les entrées de données peuvent être associé à des classes ou des concepts , ces classes et concepts sont décrites selon un mécanisme de **caractérisation des données**, en résumant les données de la classe étudiée (souvent appelé le classe cible) en termes généraux, ou **discrimination des données**, par comparaison de la classe cible avec une ou plusieurs classes comparatives (souvent appelée la en contraste Des classes),ou une composition des deux techniques. le résultat de caractérisation donnent en sortie des statistiques résumé en diagramme à barres , courbes , ou bien des règles de généralisations , alors que la discrimination donne en sortie des règles de discriminations.

1.3.2 Motifs fréquents , associations et corrélations

Les motifs fréquents sont des modèles qui apparaissent de façon assez répétitive dans les données, il existe plusieurs types tel que les ensembles d'éléments fréquent,des sous-projets ou encore des sous-structures fréquentes séquentiels et structurés de différentes formes (graphes,arbre ...), **Association** les règles d'association sont des règles exprimés en prédicat permettant de découvrir une relation entre différente variable (attributs) dans une large base de données .
correlation est une relation liant deux ou plusieurs attributs associées dans un large volume de données.

1.3.3 Classification et régression pour une analyse prédictive

La Classification est le processus de recherche d'un modèle (ou fonction) qui décrit et distingue des classes de données ou des concepts. ils sont utilisé pour prédire l'étiquette de classe d'objets pour laquelle l'étiquette de classe est inconnue

Régression régression modélise des fonctions à valeurs continue , c'est une méthode statistique très utilisée pour prédire les données manquantes ou valeurs indisponibles de données numériques plutôt que des étiquettes de classe (discrètes).

1.3.4 clustering

Le clustering est le processus de regroupement des objets (des données) selon un degré de similarité entre eux, tel que on maximise la similarité entre les membres d'un même groupe tout en minimisant la similarité entres les membres de groupes différents (dissimilarité) . ça permet une meilleure organisation de données où chaque groupe présentent une classe ,catégorie ...

1.3.5 Outlier (valeurs aberrantes)

Ce sont des objets non conformes au comportement ou au modèle général des données. Ces objets de données sont valeurs aberrantes. ces objets peuvent être vu comme du bruit ou des exceptions .

1.3.6 Ce qui fait qu'un modèle est intéressent

Il est clair que le data mining peut générer des milliers de modèles et motifs mais l'utilité des modèles restent relatif selon le besoin , à noter que les modèles sont coûteux à produire donc il faudrait s'intéresser qu'aux modèles pertinents, pour qualifier un modèle de tel adjectif il faut :

- qu'il soit facile à assimiler et comprendre par ces utilisateurs
- qu'il procure une validation de nouvelles données ou d'essai avec un certain degré de certitude
- utile et intéressent dans le sens ou il valide ce que l'utilisateur cherche à confirmer ceci donnera naissance à DES CONNAISSANCES

.

1.4 Quelles genre de technologies sont appliquées et utilisées dans le data mining ?

1.4.1 Statistique

les statistiques étudient la collection, l'analyse, l'interprétation ou l'explication et la présentation de données.UNE modèle statistique est un ensemble de fonctions mathématiques décrivant le comportement de les objets d'une classe cible en termes de variables aléatoires et leurs distributions . on peut citer la moyenne , la médian , la variance et l'écart type.

1.4.2 Machine learning (apprentissage automatique)

Apprentissage machine étudie comment les ordinateurs peuvent apprendre basé sur des données. Un domaine de recherche principal concerne les programmes informatiques qui apprennent automatiquement à reconnaître des modèles complexes et prendre des décisions intelligentes basées sur des données. on distingue deux types d'apprentissage supervisé (prédire une classe)et non supervisé (clustering par exemple) ,Semi-supervisé ,active learning

1.4.3 Système de base de données et entrepôts de données

La recherche de systèmes de bases de données se concentre sur la création, la maintenance et l'utilisation de bases de données pour les organisations et les utilisateurs . ces systèmes sont utilisé pour les raisons suivantes : les modèles de données, les langages de requête, le traitement des requêtes,méthodes d'optimisation, stockage de données et méthodes d'indexation et d'accès qui permet une meilleure exploration de données.

on note aussi l'entrepôt qui consolide les données dans un espace multidimensionnel pour former des cubes de données facilitant ainsi leurs explorations .

1.4.4 Recherche d'information

recherche d'information(IR) est la science de la recherche de documents ou d'informations dans les documents.ils peuvent être du texte ou du multimédia et peuvent résider sur le Web .La recherche d'informations suppose que les données sous recherche sont non structurées et les requêtes sont formées principalement par des mots-clés.

1.5 Quelles sont les genres d'applications ciblé par le data mining

1.5.1 Business intelligence

La technologie de l'intelligence d'entreprise (BI) fournit des informations historiques, actuelles et vues prédictives des opérations commerciales. ceci en incluant les rapports, l'analyse en ligne traitement, gestion des performances de l'entreprise, veille concurrentielle, analyse comparative et analyse prédictive de comportement des clients par exemple .

1.5.2 Web Search engine

Un Moteur de recherche Web est un serveur informatique spécialisé dans la recherche des informations sur le Web , les résultats de la recherche d'une requête utilisateur sont souvent renvoyés sous forme de liste,les types des résultats peuvent consister en des pages Web, des images et d'autres types de fichiers.

1.6 Les problèmes qui font face au data mining

Il y a beaucoup de défis en jeux de la recherche en exploration de données:

- La méthodologie minere
- Interaction avec l'utilisateur
- L'efficacité et l'évolutivité, et la gestion de diverses Types de données

Chapter 2

Connaitre vos données[1]

Dans Data Mining il y a "DATA", Ainsi nous nous devons tout d'abord d'étudier les caractéristiques des attributs et des valeurs de nos données avant de passer au pré-traitement. et ce car les données de nos Dataset collectées du monde réel contiennent du bruit, aussi elles proviennent de divers sources donc elles ne sont pas homogènes en plus de l'énorme volume qu'elles constituent. Cette étape consiste principalement en:
Trouver le type de valeur que peut prendre chaque attribut et si elles sont discrètes ou continues, Avoir une idée d'à quoi ressemble nos données et de leur distribution, Chercher quelle méthode de visualisation nous donnerai un meilleur aperçu, Mesurer le degré de similarité entre certaines données et Calculer les statistiques de base concernant chaque attribut

2.1 Type des attributs d'un Dataset

Un Dataset est un ensemble d'Objets dit données, échantillons ou tuples. Ces données sont généralement décrites par des attributs.
Ainsi dans ce qui suit nous allons plus nous intéresser aux types d'attributs existants.

2.1.1 Définition d'un attribut

Un attribut est un domaine de données, représentant une caractéristique de cette dernière. Un ensemble d'attributs aussi dit vecteur d'attribut est utilisé pour décrire un objet.
Il y a plusieurs manières de définir les types d'attributs, ceux la ne sont pas mutuellement exclusif, Le type d'un attribut est défini pas l'ensemble de ses potentielles valeurs.

2.1.2 Attributs nominaux

Nous savons d'ors et déjà que nominal signifie "relatif au nom", ainsi sa valeur est un symbole, un nom, un objet...Un attribut nominal exprime un sorte de catégorisation ou d'état de l'objet.

2.1.3 Attributs binaires

Un attribut binaire est un attribut nominal à deux états seulement: 0 et 1.

2.1.4 Attributs ordinaux

Il s'agit des attributs à valeur possédant un ordre significatif sans pour autant avoir connaissance de l'écart entre les valeurs successives.

2.1.5 Attributs numérique

Les attributs numérique sont quantitatives et représentés par des entiers ou des réels. Ils peuvent être de deux types:

Attributs à l'échelle d'intervalle

Les attributs à intervalle sont mesurés sur une échelle d'unités égales, ils suivent un ordre, leurs valeurs sont comparables et quantifiables et ne possèdent pas de vrai point de début "zéro".

Attributs avec rapport

Quant aux attributs avec rapport, ils sont numérique liés à un point d'origine, ordonnables et des rapport entre les valeurs peuvent être calculé,comme la moyenne, la médiane, le mode, ...

2.1.6 Attributs à valeurs discrètes et Continues

Un attribut dit discret prend ses valeurs dans un ensemble fini de valeur ou dans un ensemble infini mais dénombrable.
Ce type d'attribut est communément représenté (codifé) en tant qu'entier.
Dans le cas contraire, l'on parle d'attribut continu ayant pour valeur des réels, typiquement représenté par une variable à virgule flottante.

2.2 Description statistique des données

Cette partie est essentiel dans l'étude des caractéristiques: tendance centrale, variation et propagation pour la détection du bruit.

2.2.1 Mesurer la tendance centrale: moyenne, médiane et mode

Moyenne

La moyenne est l'une des mesures les plus efficace et plus couramment utilisée,
Moyenne arithmétique: soit $x_1, x_2, ..., x_n$ l'ensemble de N valeurs numériques d'un attribut **X**.

$$\text{Moyenne} = \bar{X} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Moyenne arithmétique pondérée il arrive que des poids soit associés au valeurs x_i, ainsi le calcule de la moyenne est donnée par la formule suivante:

$$\text{Moyenne} = \bar{X} = \frac{\sum_{i=1}^{N} w_i * x_i}{\sum_{i=1}^{N} w_i}$$

Note: la moyenne est sensible (influençable) par le bruit.

médiane

Soit $x_1, x_2, ..., x_n$ l'ensemble de N valeurs d'un attribut **X** selon un ordre croissant.
La médiane est alors la valeur centrale si N est impair, ou moyenne des deux valeurs moyennes sinon.

Vu la taille de N pouvant être très grande, pour les données groupées, l'on peut se baser sur l'estimation par interpolation., et ce avec la formule:

$$\text{Med} = L_l + \frac{\frac{N}{2} - (\sum_{freq})_l}{ferq_{med}} width$$

Avec: $L_l =$ la limite inférieure de l'intervalle médian, $\sum freq_l =$ somme des fréquences de tous les intervalles inférieurs à l'intervalle médian, $freq_{med} =$ fréquence de l'intervalle médian, width = la taille de l'intervalle médian.

Mode

Le mode est la valeur la plus fréquente d'un attribut, il peut exister pour un ensemble de valeurs : un mode (unimodal), deux modes (bimodal), trois modes (trimodal), à partir de trois modes on utilise aussi le terme (multimodal).
Aussi si toute les valeurs sont uniques ⇒ pas de mode.

Nous avons la relation empirique suivante:

moyenne - mode ≈ 3 * (moyenne - médiane).

Milieu de gamme

Peut aussi etre utilisé pour les attributs à valeurs numérique, comme suit:

$$\text{Milieu} = \frac{min+max}{2}$$

Note:
Moyenne = médiane = mode ⇒ distribution **symétrique** des données.
Moyenne < médiane < mode ⇒ distribution des données **asymétriques négativemen**t.
Moyenne > médiane > mode ⇒ distribution des données **asymétriques positivement**.

2.2.2 Mesure de la dispersion des données

Rang

soit $x_1, x_2, ..., x_n$ l'ensemble de N valeurs numériques d'un attribut **X**.

$$\text{Rang} = x_{MaxVal} - x_{MinVal}$$

Quartiles

Le quartile Q_1 , représente les 25% inférieurs des données ordonnées.
Le quartile Q_3 , représente les 25% supérieurs des données ordonnées.

Rang d'inter-quartile

Le Rang d'inter-quartile est la distance entre Q_1 et Q_3 est une mesure de dispersion qui donne la plage couverte par la moitié médiane des données. Elle est définie comme suit:

$$\text{IQR} = Q_3 - Q_1$$

Résumé à cinq chiffres d'un distribution et boites à moustaches

Le résumé à cinq chiffres est : Minimum, 1^{er} quartile, 3^{eme} quartile, la Médiane et Le Maximum de l'ensemble des valeurs d'un attribut (numérique).

Justement les boîtes à moustaches sont un moyen très utilisé pour visualiser une distribution se basant sur le résumé en cinq chiffres comme suit:

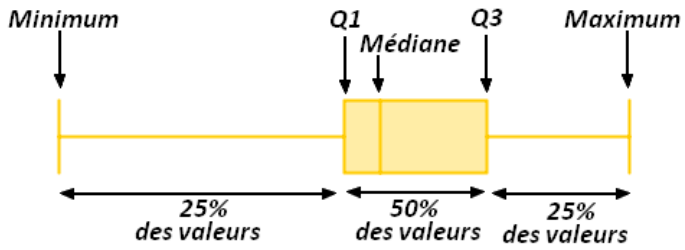

Figure 2.1: Composantes d'une boite à moustaches.

Variance et la Déviation standard

La variance et la Déviation standard sont des mesures de dispersion des données.
Une faible dérivation standard indique que les données tendent à être proches de la moyenne, Tandis qu'une dérivation standard élevée implique le contraire.

La variance de N observations, $x_1, x_2, ..., x_N$,d'un attribut numérique X est:

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

σ: est la dérivation standard.

Représentations graphiques de descriptions statistiques ds données

Tracé de quartile:Représente toutes les données, mettant en évidence aussi bien le comportement général que les valeurs inhabituelles.
Tracé Q-Q:Représente graphiquement les quantiles (Q_1, Q_2, Q_3) d'une distribution univariée par rapport aux quantiles d'une autre distribution.
Histogrammes:Représente la proportion de chaque catégorie. Lorsque les catégories ne sont pas de largeur uniforme l'on peut exploiter la surface des barres pour en tirer plus d'informations.En plus de s'adapter à tous type d'attribut.
Nuages de points:Permet un aperçu des données à deux variables, fait apparaitre les clusters et les outliers.

2.3 Visualisation des données

2.3.1 Technique de visualisation Orientée pixels

Chaque donnée de dimensions N, la visualisation se fera sur N fenêtres, une par dimension.Aussi les couleurs des pixels reflètent la valeur correspondante.

2.3.2 Techniques de visualisation par projection géométrique

Utilisée dans la compréhension de la distribution des données dans un espace multidimensionnel, citons quelques méthodes:
-Matrices de nuages de points. -Nuages de points. -Vues de la section -Hyperslice -Coordonnée parallèle ...etc

2.3.3 Techniques de visualisation basées sur des icônes

cette technique utilise de petites icônes pour représenter des images multidimensionnelles. Citons deux méthodes des plus populaires: *Chernoff Faces* et *Stick Figures*.

2.3.4 Techniques de visualisation hiérarchique

Il s'agit d'un partitionnement hiérarchique en sous-espaces, parmi les méthodes connues: "Worlds-within-Worlds","Tree-map", ...

2.3.5 Visualiser des données complexes et des relations

Permet de visualiser des données non numériques tel du texte (données complexe) et des réseaux sociaux (relations). Comme un nuage de tags est une visualisation des statistiques des tags générés par l'utilisateur.

2.4 Mesure de la similarité et de la disparité des données

Mesure numérique du degré de ressemblance de deux objets, généralement ayant des valeurs entre o et 1.

2.4.1 Matrice de données versus matrice de disparité

Matrice de données: stockant les valeurs des données. sous la forme d'une table relationnelle (matrice) X objets x N attributs.

matrice de disparité: renferme la distance entre les objets d'un ligne et ceux d'une colonne. (matrice X Objets x X Objets).

2.4.2 Mesures de proximité pour les attributs nominaux

Méthode 1:Simple correspondance: Matrice de disparité, avec:

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

Avec : m: nombre de correspondance, p: nombre total de variables.
Méthode 2: utiliser un grand nombre d'attributs binaires:

2.4.3 Mesures de proximité pour les attributs binaires

Il existe:
Mesure de distance pour les variables binaires symétriques.
Mesure de distance pour les variables binaires asymétriques
Coefficient de Jaccard (mesure de similarité pour les variables binaires asymétriques).

2.4.4 Dissimilarité des données numériques: distance de Minkowski

Cette mesure comprend la distance Euclidienne et la distance de Manhattan, sa formule :

$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + ... + |x_{in} - x_{jn}|^h}$$

Avec: n = nombre d'attributs, h = nombre réel tel que h $\geq$ 1.

2.4.5 Mesures de proximité pour les attributs ordinaux

les étapes suivantes sont nécessaires:
-Remplace x_{if} par son rang.($r_{if} \in 1, ..., M_f$).
-Normalisez les variable sur [0, 1] en remplaçant le i-ème objet de la f-ème variable par:

$$z_{if} = \frac{r_{[if]}-1}{M_{if}-1}$$

-Mesurer la distance.

2.4.6 Dissemblance des attributs de types mélangés

Notre Dataset peut contenir tous les types d'attributs, calcule de la distance selon la formule:

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} dij^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

Avec: f est binaire ou nominal, $dij^{(f)} = 0$ si $x_{if} = x_{jf}$ ou $dij^{(f)} = 1$ sinon f est numérique: utilise la distance normalisée.

2.4.7 Similitude cosinus

C'est une mesure de la similarité qui peut être utilisée pour comparer des documents, de très grands nombres d'attributs, comme la fréquence des mot dans les documents (ou plusieurs mots peuvent avoir une fréquence = 0).
Soit x et y deux vecteurs de comparaison

$$sim(x,y) = \frac{x \bullet y}{||x|| * ||y||}$$

Avec: $||x||$ est la norme Euclidienne du vecteur $x = (x_1, x_2, ..., x_n)$

Chapter 3

Pré-traitement des données[1]

3.1 Qualité des données : Pourquoi faire du prétraitement sur nos données ?

On dira que les données sont de qualité si elles répondent aux exigences suivantes : l'exactitude, exhaustivité, complétude, cohérence, actualité (évolutif dans le temps), crédibilité et interprétabilité.

Mais avoir des données de qualité n'est pas chose facile , dans le monde réel beaucoup de facteurs rentrent en jeux pour compromettre la qualité des données tel que :

- valeurs manquantes (donnée incomplète) dans les tuples de données
- valeurs aberrantes (date de naissance 1700 par exemple "bruité")
- inconsistance ou encore des valeurs dupliquées .

Si alors on plonge directement dans l'exploration de nos données sans prêter attention à ces inconvenants là, la qualité des motifs ou modèles sera affectés directement ce qui causera un manque de crédibilité et de confiance dans ces derniers qui risquent être erronés.

3.2 Data Cleaning

L'une des principales approches de prétraitement des données , qui essaye de proposer des solutions aux différents problèmes citées plus .

3.2.1 Valeurs manquantes

Il existe diverse méthodes pour gérer les valeurs manquantes des données parmi elles :

Ignorer le tuple

cela consiste à renoncer à l'utilisation du tuple alors que ce dernier peut s'avérer intéressent par la suite, cette solution n'est efficace si la majorité de nos données souffrent d'un manque de valeurs d'un attribut ou deux .

Remplir les valeurs manquantes manuellement

Si la taille de nos données est conséquente cette méthode devient infaisable (elle prend énormément de temps) .

Utilisation d'une valeur globale pour remplir les valeurs manquantes

son inconvénient majeur est que lors de la fouille ou l'exploration ce genre de valeurs globales peuvent être prise pour des concepts intéressent .

Utilisation d'une mesure de la tendance centrale de l'attribut (par exemple, la moyenne ou la médiane)

Utilisez l'attribut moyenne ou médiane pour tous les échantillons appartenant à la même classe que le tuple donné

Utilisation de la valeur la plus probable

ceci peut être fait grâce à la fonction de régression , arbre de décision ou encore l'algorithme naïf de bayes,c'est généralement la technique la plus utilisée .

3.2.2 Les données bruitées ou erronées (Noisy data)

Les données bruitées sont des données qui manquent d'incohérence comme des valeurs aberrantes ,plusieurs techniques ont été mise au point pour faire face à ce genre de problème ,on citera:

Binning (nettoyer en enlevant des valeurs)

Binning se repose principalement sur le trie des valeurs des données puis il effectuera un partitionnement uniforme de façon à voir la même taille pour chaque partition ,ce qui permettra de consulter les voisins des valeurs erronées (une recherche locale) , le binning choisira par la suite de remplacer les valeurs aberrantes par la moyenne de la partition ou la médiane ou encore par le min si la valeur est plus proche du celui ci sinon par le max de l'intervalle de la partition (le plus proche voisin).

Régression

Afin de remplacer les valeurs erronés on peut faire appelle à une fonction de régression linéaire qui inclut deux variables (attributs) on donnera alors une valeur d'attribut pour prédire l'autre.

$$F(attribut_1) = attribut_2$$

ou bien une fonction de régression de dimension trois ou plus ou plusieurs attributs sont mise en jeux .

$$F(attribut_1, attribut_2, ..., attribut_n) = attribut_j$$

Analyse des outlier

en utilisant le clustering

3.2.3 Data Cleaning comme un processus

On peut résumer le nettoyage de données en des étapes élémentaires qui sont les suivantes:
Détection de divergence :

- Utilisez des métadonnées .
- Vérifiez la surcharge du champ.
- Vérifiez la règle d'unicité, la règle consécutive et la règle null.
- Utilisez des outils commerciaux.(data scrubbing (nettoyage) ,Data auditing(detecter des relations o corrélations)

Migration de données et intégration:

- Outils de migration de données
- Outils ETL (Extraction / Transformation / Loading)

Intégration des deux processus: Itératif et interactif

3.3 Intégration des données

3.3.1 Problème d'identification d'entité

ce problème survient lorsque on à faire à des données collecter de plusieurs sources, l'identification d'entité consiste à retrouver des attributs qui font référence à la même entité par exemple l'attribut id_consommateur ou nb_consommateur mais aussi la détection de différente structure d'un même attribut du diverse sources , généralement on utilise les méta données pour remédier à ce problème

3.3.2 Analyse de redondance et de corrélation

Lorsqu'un attribut peut être déduit à partir d'un autre attribut ou plus on dira qu'il est redondant, l'analyse de corrélation permet de détecter la redondance on citera :

X^2 corrélation pour données nominales

Pour les données nominales, une relation de corrélation entre deux attributs, A et B , peut être découvert par un X2(chi-carré) test, Hypothèse : les deux distributions sont indépendantes Les cellules (matrice r colonnes c lignes) qui contribuent le plus à la valeur X2 sont celles dont le nombre réel (observé) est très différent du nombre attendu . Plus la mesure X2 est grande, plus les variables sont susceptibles d'être liées.

$$X^2 = \sum_{i=1}^{C} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$e_{ij} = \frac{count(A = a_i) * count(B = b_j)}{n}$$

c : valeurs distinctes que peut prendre l'attribut A
r :valeurs distinctes que peut prendre l'attribut B
o_{ij}: fréquence observé de l'événement commun $A = a_i, B = b_j$
e_{ij}: fréquence attendue (prévisible) de l'événement commun $A = a_i, B = b_j$
n: nombre de tuples totale
count(a_i): nombre de tuple ayant la valeur a_i pour A
count(b_j):nombre de tuple ayant la valeur b_j pour B

Coefficient de corrélation pour les données numériques

coefficient de corrélation de pearson :

$$rA, B = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i, b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B}$$

$\bar{A}$:La moyenne des valeurs de l'attribut A.
$\bar{B}$:La moyenne des valeurs de l'attribut B.
σ_A: l'écart type de A
σ_B: l'écart type de B
$\sum_{i=1}^{n}(a_i, b_i)$: produit vectoriel
r A,B est une mesure entre -1 et 1 si la valeur est positive on dira que A ,B sont positivement corrélé donc ils augmentent de façon proportionnelle, plus la valeur est proche de 1 plus ils sont corrélé , par contre si la valeur est négative on dira que A,B sont inversement proportionnelle . à noter que si r A,B $= 0$, A et B sont indépendants

Covariance des données numériques

Dans la théorie des probabilités et les statistiques, la corrélation et la covariance sont deux mesures similaires.

$$E(A) = \bar{A} = \frac{\sum_{i=1}^{n} a_i}{n}$$
$$E(B) = \bar{B} = \frac{\sum_{i=1}^{n} b_i}{n}$$

La covariance est défini comme :

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$
$$rA, B = \frac{Cov(A, B)}{\sigma_A\sigma_B}$$
$$Cov(A, B) = E(A.B) - \bar{A}\bar{B}$$

Comme la corrélation numérique ,A,B sont dites indépendant alors leurs covariance est nulle , positivement proportionnelle si la covariance est positive et inversement proportionnelle si covariance est négative.

3.3.3 Duplication de tuples

c'est lorsqu'il y a deux ou plusieurs tuples identiques pour une donnée en cas de saisie de données unique.

3.3.4 Détection et résolution de conflits de valeurs de données

L'intégration de données implique également la détection et résolution de conflits de valeurs de données par exemple pour une même entité du monde réel, les valeurs d'attribut provenant de sources différentes peuvent varier. Cela peut être dû à des différences de représentation, de mise à l'échelle ou d'encodage. Par exemple, un poids (attribut) peut être stocké en unités métriques (kg) dans un système et impériale britannique dans un autre (pound).

3.4 Réduction de données

Des techniques de réduction des données permettent d'obtenir une représentation réduite du Dataset, tout en maintenant l'intégrité du document d'origine. Ainsi l'extraction de données devrait être plus efficace tout en produisant des résultats d'analyse quasi identiques.

3.4.1 Les stratégies de réduction de données

Les stratégies de réduction des données incluent:
La réduction de la dimensionnalité: permet d'éliminer les attributs non pertinents et de réduire le bruit, parmi les techniques: les transformées en ondelettes et l'analyse en composantes principales.
La réduction de la numérotation, elle remplacer le volume de données d'origine par une alternative, citons: les modèles de régression et log-linéaires, Histogrammes, regroupement, échantillonnage, Agrégation de cube de données.
La compression des données:cette stratégie permet d'obtenir une représentation réduite ou <<compressée>> des données d'origine.

3.4.2 Transformée en ondelettes

C'est une technique de traitement du signal linéaire étroitement liée à la transformée de Fourier discrète qui, appliqué à un vecteur de données X, le transforme en un vecteur différent X' de coefficients d'ondelettes. En ne stockant qu'une petite fraction du plus fort des coefficients.

3.4.3 Analyse en composantes principales

Cette analyse permet de trouvez une projection qui englobe la plus grande quantité de variation dans les données.

3.4.4 Sélection de sous-attributs

Cette stratégie à fait ses preuves pour l'élimination d'attributs redondants et non pertinents. Quelques méthodes typiques de sélection d'attributs par heuristiques: Sélection pas à pas, élimination progressive, Combinaison de sélection en avant et d'élimination en arrière et Induction par arbre de décision.

3.4.5 Modèles de régression et log-linéaires: Réduction de donnée paramétriques

Des modèles de régression et log-linéaires fournissent une approximation des données fournies.
Régression linéaire: Données modélisées pour correspondre à une ligne droite.
La régression linéaire multiple:il s'agit d'une extension de la précédente, elle permet à une variable de réponse y, d'être modélisé comme une fonction linéaire de deux variables prédictives ou plus.
Modèles log-linéaires:Approximation des distributions de probabilité multidimensionnelles discrètes.

3.4.6 Histogrammes

Représente la distribution des données, Il divise les données en compartiments et garde la moyenne pour chacun d'eux.
Deux règles de partitionnement sont possibles: Largeur égale ou fréquence égale.
Utiles pour approximer les données éparses et denses, très asymétriques et uniformes.

3.4.7 Clusterisation

A pour but de partitionner les données en clusters en fonction de la similarité.

3.4.8 Échantillonnage

il permet à un grand ensemble de données d'être représenté par un échantillon de données aléatoire beaucoup plus petit.
Voici quelques type d'échantillonnage: Échantillonnage aléatoire simple, Échantillonnage sans remplacement, Échantillonnage avec remplacement, Échantillonnage stratifié.

3.4.9 Agrégation de cube de données

Il utilise le niveau le plus bas d'un cube de données (cuboïde de base), en utilisant la plus petite représentation qui soit suffisante pour résoudre la tâche.

3.5 Transformation et discrétisation des données

3.5.1 Vue d'ensemble des stratégies de transformation de données

Lissage: Il permet d'éliminer le bruit des données. Les techniques incluent le binning, régression et clustering.
Construction d'attribut:de nouveaux attributs sont ajoutés à partir de l'ensemble d'attributs donné pour aider le processus d'exploration.
Agrégation:opération de résumé ou fonction d'agrégation, généralement utilisée dans la construction du cube de données.
Normalisation:mises à l'échelle des valeurs des attributs de manière à réduire la plage de données.
Discrétisation: lorsque les valeurs brutes d'un attribut numérique sont remplacées par étiquettes

d'intervalle.
Génération de hiérarchie de concepts pour les données nominales:pour des attributs qui peuvent être généralisé.

3.5.2 Transformation de données par normalisation

La normalisation des données tente d'attribuer un poids égal à tous les attributs, pour n'en favoriser aucun. Il existe de nombreuses méthodes de normalisation des données. Nous étudierons la normalisation min-max, z-score et par mise à l'échelle décimale. Pour ce qui suit, Soit A un attribut numérique avec n valeurs observées, $v_1, v_2, ..., v_n$.
normalisation min-max:effectue une transformation linéaire sur les données d'origine.
la normalisation d'un v_i de $[min_A, max_A]$ dans $[new\_min_A, new\_max_A]$

$$v'_i = \frac{v_i - min_A}{max_A - min_A} * (new\_max_A - new\_min_A) + new\_min_A$$

normalisation z-score:les valeurs pour un attribut A, sont normalisés en fonction de la moyenne et l'écart type.

$$v'_i = \frac{v_i - \bar{A}}{\sigma_A}$$

normalisation par mise à l'échelle décimale:normalise en déplaçant le point décimal des valeurs d'attribut A. (j est le plus petit entier tel que $max(|v_i|) < 1$)

$$v'_i = \frac{v_i}{10^j}$$

3.5.3 Discrétisation par binning

Le binning est une technique de fractionnement par le haut.

3.5.4 Discrétisation par analyse d'histogramme

Un histogramme partitionne les valeurs d'un attribut A, en plages disjointes appelées compartiments.

3.5.5 Discrétisation par cluster, arbre de décision,et analyses de corrélation

Il s'agit dans l'ordre de :"fusion non supervisée, scission ascendante ou ascendante","scission supervisée et descendante" et "fusion non supervisée et non supervisée".

3.5.6 Génération de hiérarchie de concepts pour les données nominales

Organise les valeurs d'attributs de manière hiérarchique et est généralement associé à chaque dimension d'un entrepôt de données.citons 4 méthodes:

- Spécification d'un classement partiel des attributs explicitement au niveau du schéma par utilisateurs ou experts.
- Spécification d'une partie d'une hiérarchie par regroupement explicite de données.
- Spécification de seulement un ensemble partiel d'attribut.
- Génération automatique de hiérarchies par l'analyse du nombre de valeurs distinctes.

Bibliographie

[1]Jiawei Han, Micheline Kamber, Jian Pei."Data-Mining Concepts and Techniques", 3rd Edition Morgan Kaufmann (2011).